# Review of Physics in Technology

Contents


Editorial 1

Articles

ESRO and its programme 2
*H Bondi and J Ortner*

Molecular engineering in the design of microwave ferrimagnetic garnets 9
*A S Hudson*

Recent developments in conventional soft magnetic materials 27
*J E Thompson*

Back to the old Trend: arguments and background to the Green Paper 47
*C L Boltz*

Book Reviews 58

Forthcoming Courses and Conferences 63

Sources of Reviews 64



The Institute of Physics and The Physical

# Review of Physics in Technology

**Chairman of the Editorial Advisory Panel**
Professor L R G Treloar

**Editorial Advisory Panel**
Dr K W Andrews
M F Avery
C G Cannon
Dr P Dean
Dr J S Halliday
Professor D H Parkinson
Dr H O Puls
Dr M J Stowell
Professor K W Keohane
Professor J Ring
Professor J M Ziman

**Assistant Editor**
P A Tulett

**Editorial and Advertisements Departments**
The Institute of Physics and
The Physical Society
1 Lowther Gardens
Prince Consort Road
London SW
England

Manuscripts and correspondence relating to the editorial and advertisement content should be sent to the address given above.

**Subscription Information**
The *Review of Physics in Technology* is published three times a year by The Institute of Physics and The Physical Society, 47 Belgrave Square, London SW1, England in association with the American Institute of Physics, and is printed in Great Britain by Adlard & Son Ltd, Dorking, Surrey. Remailed and 2nd class postage paid at Prince and Lemon Streets, Pennsylvania 17604, U S A.

The public subscription rates for all countries, except the United States, Canada and Mexico, is £3 and orders should be sent to the Circulation Department, The Institute of Physics and The Physical Society, 47 Belgrave Square, London SW1, England.

Members of the Institute and Society may subscribe to the journal at the special rate of £2. This special rate is not applicable to orders placed through the American Institute of Physics. Members in North America may remit in U S dollars to the Chase Manhattan Bank, 1 Chase Manhattan Plaza, New York 10015, U S A, for transfer to the account of the Institute of Physics and The Physical Society at the Chase Manhattan Bank, London, or they may remit in U S dollars or sterling to the Circulation Department, The Institute of Physics and The Physical Society, 47 Belgrave Square, London SW1.

The rate for North America by airfreight is $7.20. Orders from North American countries (except from members, and for back issues of the current volume and single copies) should be sent to the American Institute of Physics, 335 East 45th Street, New York 10017, U S A.

The single issue price is £1. All orders for single issues should be sent to the Circulation Department, The Institute of Physics and The Physical Society, 47 Belgrave Square, London SW1, England.

Copyright © 1970 by The Institute of Physics and The Physical Society and individual contributors. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without permission of the copyright owner.

# Review of Physics in Technology

## Contents

**Articles**

Direct current electroluminescence in ZnS 65
*A Vecht and N J Werring*

The physics of Schottky barriers 81
*E H Rhoderick*

A review of progress in gas lubrication 96
*J W Powell*

Nobody breathes fresh air: Facts and arguments about air pollution 130
*C H Boltz*

**Book Reviews** 143

**Forthcoming Courses and Conferences** 144

**Sources of Reviews** 144

**The Institute of Physics and The Physical Society**

# Review of Physics in Technology

## Contents

**Articles**

The hardness of solids 145
*D Tabor*

Optical spectroscopy in the frequency range $1–10^8$ Hz 180
*E R Pike*

Information: The essential commodity: Thoughts based on the Paris seminar of March 1970 195
*C L Boltz*

**Book Reviews** 205

**Book Notices** 208

**The Institute of Physics and The Physical Society**

# Editorial

With the increasing rate of expansion of scientific publications the individual physicist is faced with the increasingly difficult problem of trying to keep in touch with developments in the various branches of the subject in which he is interested but for which he cannot possibly find the time to study the original literature in detail. The object of a review journal is to go some way towards solving this problem, by providing the reader with a coherent presentation of developments in a particular field in a more readable form. The needs of those working in the basic physical sciences are already catered for by a number of review journals, which include Reports on Progress in Physics, first published in 1934, but the needs of the physicist in the industrial environment, who is concerned primarily with the applications of physics in the field of technology, have up to the present received less attention. It is for this reason that this new journal, *Review of Physics in Technology*, is being launched.

It is intended that the style and content of the journal will be such as to attract individual subscribers. In pursuance of this aim it is felt desirable to limit both the amount of material in each issue and the number of issues per annum to a level which can be conveniently assimilated by the individual reader.

The backbone of the journal will be a planned series of review articles (approximately three per issue) written primarily for those who are not specialists in the particular subjects covered. Future issues will also contain regular features or shorter articles dealing with various aspects of physics and its technological applications. In addition, correspondence from readers, book reviews, notices of courses and other matters of current interest will be included.

The success of this venture will depend very much on the degree to which it satisfies the needs of our readers. I would therefore like to take this opportunity of inviting readers to write to the Editor on any matters concerning the journal and its contents on which they would care to express an opinion. It is hoped to publish correspondence in the next issue, which will appear in August.

L R G Treloar
Chairman of the
Editorial Advisory Panel

H. BONDI and J. ORTNER
ESRO, Paris

# ESRO and its programme

**Abstract.** The aims, structure and achievements of the European Space Research Organisation are reviewed.

## 1. Introduction

Space research is the task for which ESRO (European Space Research Organisation) was founded in 1964, but space research is not a single field of science like, say, molecular biology or high-energy physics, but a coalition of many different sciences united solely by the means adopted, the use of space platforms. Thus ionospheric physics, aeronomy, the study of solar–terrestrial relations, plasma studies, magnetospheric research, astronomy in very low frequencies, in the infra-red, ultra-violet, X-ray and $\gamma$-ray portions of the spectrum, planetary and lunar studies and exobiology, all come under the umbrella of space research. Space platforms are expensive to build, launch, and operate, and so this is 'big science' with a vengeance. Because of the high cost, both of the product and of the facilities required, ten European countries,† egged on by their scientists, formed ESRO in the Spring of 1964 after two years of preparation in the COPERS organization. It was then decided that ESRO would employ both long-life space platforms (satellites and interplanetary probes) and short-term ones (sounding rockets). There are a number of balancing acts to be performed in ESRO. The resources available have to be divided between sounding rockets and satellites, between investment in facilities and the advancement of European technology (which is one of ESRO's main aims) on the one hand, and operations (completion of sounding rockets and satellites) on the other, between the interests of the big member states (who have national programmes in space and are mainly interested in ESRO for projects bigger than can be funded nationally) and the smaller member countries, for whom ESRO is the only avenue to space, between the demands of the different sciences, between the political pressures of different governments behind whom stand industries, scientists and populations with differing aims and tastes. Overshadowing all is the necessarily high cost of space work. This is due to a variety of factors. First, there is the need for spacecraft with their many sub-systems (telemetry, telecommand capability, attitude control, power supply and, above all, the experiments themselves) to function after the severe vibration of a launch in the utterly strange and demanding environment of space. In a sounding rocket, unless everything works for the few minutes of actual space exposure, the project is a failure. In a satellite, reliable operation without availability of a mechanic imposes the need to ensure utter reliability through design and fabrication methods. In all cases, considerable numbers of high-grade engineering staff are required for the design and the most severe critical reviews of the design, for devising fabrication safeguards and testing procedures. Above all, the full compatibility of all the different pieces and sub-systems must be assured by a most elaborate form of management. All this requires a lot of time of a lot of highly skilled men, and this costs a lot of money. The launch vehicles themselves are of considerable expense. After the launch, the command of the systems and the determination of position and attitude, and the recovery of the scientific data (which after all is *the* end product) all demand expensive

† They are Sweden, Denmark, The Federal Republic of Germany, The Netherlands, Belgium, The United Kingdom, France, Switzerland, Italy and Spain. Austria and Norway, though initially interested, did not join ESRO but have observer status.

installations and make considerable demands on staff, in quantity and quality. Again there are differences. Sounding rockets demand ranges with precision radar, extensive telemetry, helicopters for recovery, and large areas temporarily sealed-off for safety reasons, while satellites require a world-wide chain of telemetry stations, tracking capability, and an operations centre to decide on necessary telecommands. In the end, massive computer operations are required to disentangle the data and link them to attitude, position and time information.

In ESRO there is a divided responsibility, ESRO itself funding and managing the spacecraft, launch and post-launch operations, while the member states themselves have funded the individual experiments.

To give some figures, the ESRO I satellite (launched in 1968) cost nearly £ 3m to produce, with the experiments adding perhaps £ 0·6m; the launch (provided free of charge by NASA) had a real cost of nearly £ 1m, while the post-launch operations (including an appropriate slice of computer rental) added perhaps £ 0·5m. This was a satellite of modest size but considerable complexity. ESRO's first project of really 'continental' size, TD1 (due for launch in 1972) will involve spacecraft costs of at least £ 15m, with the experiments adding perhaps £ 3·5m, a launching cost of £ 2·5m and operational costs of around £ 1m.

The costs of ESRO are shared amongst member countries according to gross national product (with a 25% upper limit attained by the United Kingdom in the early days). In 1969, the United Kingdom's share of a £ 21m expenditure was about 23%.

The level of the costs is in part determined by the fact that ESRO's task is not only the furtherance of Europe's space science but equally the advancement of Europe's technology. Thus reliance on U.S. hardware (often cheaper than the European equivalent) is kept to a minimum, and ESRO acts as design authority through an industrial firm appointed as prime contractor, with subcontractors widely spread through Europe.

## 2. Structure

ESRO is governed by a Council in which each member state is represented (generally by one scientist and one civil servant) and has one vote. The Council is the ultimate authority on all matters. The organization itself is headed by the directorate, consisting of the Director General, the Director of Programmes and Planning, the Director of Administration (all three with their seat at the head office in Paris), the Director of ESTEC (European Space Research and Technology Centre), ESRO's large technical facility responsible for design and testing at Noordwijk in the Netherlands, and the Director of Space Operations who is responsible for the Operations Centre in Darmstadt (Germany), the rocket range at Kiruna (Northern Sweden), rocket operations at non-ESRO ranges (the chief ones used are in Italy, Norway and Australia) and telemetry stations in Belgium, Alaska, the Falkland Islands and Spitzbergen.

An elaborate structure exists for the selection of experiments to be flown. The Directorate of Programmes and Planning is responsible for the submission to Council (through its Scientific and Technical Committee) of mission specifications and payload composition, and also for gathering the necessary scientific advice, which is obtained through the Launching Programmes Advisory Committee (consisting of five eminent scientists serving for three years), which itself is advised by specialist groups (the quaintly named *ad hoc* groups) in the various fields. Throughout the planning period there must be the closest interplay between scientific desirabilities and technical and economic possibilities. The process is necessarily long and must be seen to lead openly to a just and fair result.

This is the case of programme making, but mention should be made of ESRO's other missions, which are on the ground rather than in space. ESRO has a small research institute, ESRIN (European Space Research Institute), at Frascati in Italy, concerned with plasma physics. Next there is an educational programme, awarding fellowships tenable at suitable European and American institutions, conducting summer schools and colloquia, etc. Finally, there is a documentation service, making available to research institutions and industry data on space hardware, software, results etc., from the complete U.S.A. and European information available to ESRO on an interchange basis with NASA.

### 3. The first four satellites

ESRO's first three satellites were launched by NASA in 1968, two of them on a co-operative basis (i.e. NASA meeting the launch costs), the third on a cost reimbursement basis.

ESRO II ('Iris'), launched on the 17th of May 1968, is in a low orbit at an inclination of 98°, i.e. a retrograde near polar orbit. Such an orbit follows the terminator for many months, during which the satellite is continually in sunlight. With a low orbit, the satellite passes over telemetry stations only for a small fraction of each orbit (which takes some 90 minutes), each station pass lasting around 10 minutes. To get all the data down, a tape recorder is necessary which on command plays back at high speed all the data received since the previous pass, so that they can all be recorded at the telemetry station. The mechanical complexity of tape recorders gives them all a limited life, and the ESRO II tape recorder broke down after 6½ months service, a slightly longer life than the space average of 5½ months. The satellite data are of such interest, however, that even after the breakdown, through the use not only of ESRO's but of other telemetry stations, a substantial fraction of data is recovered. The experiments on ESRO II are:

| | |
|---|---|
| Monitor of energetic particle flux<br>Solar and Van Allen belt protons<br>Solar and galactic α-particles and protons | Imperial College, London (H. Elliot) |
| Primary cosmic-ray electrons | University of Leeds (P. Marsden) |
| Hard solar X-rays | University College London (R. L. F. Boyd) |
| Soft solar X-rays | Space Research Laboratory, Utrecht (C. de Jager) |
| Flux and energy spectrum of solar and galactic particles | Centre d'Etudes Nucléaires de Saclay, France (J. Labeyrie). |

Though publication of data necessarily occurs some time after recovery of data, preliminary results for Iris show that all experiments have yielded good results, some of which have confirmed previous space observations and some of which have indicated that further confirmatory measurements are necessary. Particularly new and topical are the following.

(i) *High-energy primary electrons*

The electron experiment of the University of Leeds was the first satellite measurement to be made of primary electrons with energies in the range 5–20 GeV. The preliminary results indicate that previous values, obtained from the extrapolation of balloon measurements, produced fluxes too low by a factor of 5. It is clear that further work is necessary before it is possible to talk about time variations in flux or about the astrophysical implications.

(ii) *Solar X-rays*

The experiments of University College London, and Utrecht, show that the time profiles for different energy bands are very different. While this is not the very first such observation, the subject is still so new that the results are a significant contribution.

(iii) *Irregularities in the arrival of solar particles at the polar caps*

The results of the Imperial College and Saclay experiments, especially when taken in conjunction with the HEOS (Highly Eccentric Orbit Satellite) measurements, show that, contrary to what was believed a few years ago, the particles do not arrive uniformly over the polar caps. This problem requires much more observation and a significant contribution is being made, especially since the Americans have had relatively few long-lived polar satellites.

Next in line was ESRO IA, launched on the 3rd of October 1968, also into a low polar orbit, but with a magnetic attitude control which makes the satellite align with the magnetic line of force so that it has to tumble over each Pole. Again, a tape recorder was necessary, which functioned for almost 7 months, and again data recovery continued. Because of the interest

not only in continuing the work but in obtaining correlations between two different locations in space, an identical model (ESRO IB) was launched on the 1st of October 1969. This is ESRO's fourth satellite. The experiments carried on each of the two satellites are:

| | |
|---|---|
| Absolute auroral luminosity | Norwegian Institute of Cosmic Physics, Oslo (A. Omholt) |
| Electron temperature and density<br>Positive ion composition and density | University College London (R. L. F. Boyd) |
| Electron flux and energy spectra in the range 50–450 keV<br>Proton energy spectra between 5 and 30 MeV | Radio and Space Research Station (R. Dalziel) |
| Low-energy electrons and protons in the range 1–13 keV | Kiruna Geophysical Observatory (B. Hultqvist) |
| Auroral proton energy spectra in the range 0·1–6 MeV | University of Bergen (B. Trumpy) and Danish Space Research Institute, Lyngby (B. Peters) |
| Monitor of energetic particle flux | Norwegian Defence Research Establishment, Kjeller (B. Landmark) and Danish Space Research Institute, Lyngby (B. Peters). |

All experiments worked perfectly. These satellites are stabilized along the magnetic field by a bar magnet so that, over the North Pole, particle experiments look up and across the magnetic field, while experiments examining auroral light look down at the earth.

The experiments have confirmed many of the results obtained by the similar American Injun satellites. However, the particle energy ranges are more adequately covered on ESRO I, and a solar proton experiment is carried. In addition, the spatial resolution of the measurements of auroral features is superior to all previous experiments elsewhere. Of particular interest is the particle flux across the polar cap.

The proton flux can be monitored in conjunction with measurements of the trapping boundary for the electron belts, and of the position of the auroral particle dumping into the atmosphere. It is then possible to make certain reasonable guesses about the configuration of the magnetosphere and how low-energy solar particles can make their way through the magnetic field to reach the earth. The picture is far from clear, and again the new ESRO results are making a considerable contribution.

ESRO IB was intentionally injected into a low (short life) orbit to probe the F-layer of the ionosphere, an orbit which in fact turned out to be even a little lower than planned, giving it only a two-month life to re-entry.† Intensive data acquisition has been organized to compensate for the early breakdown of the tape recorder. The correlation with data from ESRO IA and HEOS is particularly valuable.

ESRO's third venture was HEOS I, launched on the 5th of December 1968 into a highly eccentric orbit, penetrating interplanetary space to a distance of some 33 earth radii but approaching the Earth closely for a short part of its 5-day orbit. Since the satellite is so distant during the most interesting part of its orbit, it is visible almost all the time from some telemetry station and does not require a tape recorder. On the other hand, telemetry is made difficult by the large distance, and so the data rate is kept low. The experiments are:

| | |
|---|---|
| Barium cloud (electric fields) | Max-Planck Institute for Extra-Terrestrial Physics, Garching (R. Lüst) |
| Magnetic fields<br>High-energy cosmic-ray anisotropy<br>Low-energy solar protons | Imperial College, London (H. Elliot) |
| Solar wind | Universities of Florence, Rome and Brussels (A. Bonetti, G. Pizzella and R. Coutrez) |

† This occurred on 23rd November 1969.

| | |
|---|---|
| Flux and spectrum of cosmic rays | C.E.N., Saclay (J. Labeyrie) |
| Primary cosmic-ray electrons | C.E.N., Saclay (J. Labeyrie) and University of Milan (C. Dilworth-Occhialini). |

All these experiments have worked well and given valuable results.

The orbital configuration of HEOS means that the most significant results are obtained in interplanetary space and at the boundary of the magnetosphere when both are under the influence of the sun at its maximum activity.

Some very important results have already been obtained from HEOS and many others can be expected when data analysis is complete. The barium cloud was released on the 18th of March 1969 when the spacecraft was at an altitude of 75 000 km. The cloud was observed from ground stations for a period of some 25 minutes. Analysis of the motion striations and expansion of the cloud has yielded fundamental information on the dynamics of the expanding plasma, including some of its instabilities under the influence of magnetic and electric fields. The magnetometer was sampling the ambient magnetic field at a high rate at the time of the cloud's release, and the observed modulation of the field has yielded additional data on the expansion process.

The solar event of the 25th of February 1969 has received considerable attention by the experimenters. Correlation studies between the solar particle experiments on HEOS and on ESRO II, and experiments flown on sounding rockets from Andenes, Norway, have shown that on this occasion the magnetosphere was open to the solar particles down to very low altitudes. The propagation of particles of solar origin through the magnetosphere is one of the outstanding problems of magnetospheric physics. The cosmic-ray electron experiment has observed high-energy electrons of solar origin during flare events, followed some days later by what appears as an 'echo', as though the electrons were reflected from some distant boundary. These observations are unique, and possible mechanisms are under study.

The magnetometer, low-energy and high-energy proton telescopes have made observations on the anisotropy of particle fluxes correlated with the magnetic field orientation. A detailed study should reveal information on the modes of propagation of charged particles through interplanetary space. The solar wind experiment has observed the arrival at the spacecraft of shock waves in interplanetary space and has studied in detail the structure of the Earth's bow shock and magnetosphere boundary. Once again, correlation with magnetometer measurements should yield fundamentally new information.

From the foregoing it can be seen how one experiment can yield valuable information which may be correlated with that from others to give a much broader picture of the physical processes under investigation. A further step has already been taken to correlate HEOS measurements with those obtained by other spacecraft, and the initial results, although preliminary, look very encouraging.

These four satellites demonstrate ESRO's capability in managing such programmes of considerable complexity. Each of the four satellites is more sophisticated than any produced so far in Europe, though U.K.3 comes close.

## 4. The early 1970's

So far as future satellites are concerned, ESRO's biggest project is TD1. Begun in 1967 as one of a pair of similar spacecraft (TD1 and TD2), the programme ran into troubles of cost escalation and management, and in mid-1968 it was decided that only TD1 could go ahead. This vast satellite (weighing nearly $\frac{1}{4}$ a ton) with corresponding complexity, particularly in attitude control, carries seven experiments, some of them individually technically as difficult as the whole of one of the earlier satellites. The launch, by NASA, is due for Spring, 1972. The experiments are:

| | |
|---|---|
| Multicolour celestial scanning in the ultraviolet (1000–3000 Å) and in the infra-red | Institut d'Astrophysique, Liège (L. Houziaux, L. Delbouille and A. Monfils) and Royal Observatory, Edinburgh (H. E. Butler) |

| | |
|---|---|
| UV stellar spectrometry (2000–3000 Å) | Space Research Laboratory, Utrecht (C. de Jager) |
| Solar x-rays (20–700 keV) | |
| Spectrometry of primary charged particles | C.E.N., Saclay (J. Labeyrie) |
| Spectrometry of celestial x-rays (2–30 keV) | |
| Solar $\gamma$-rays (50–500 MeV) | University of Milan (G. Occhialini) |
| Celestial $\gamma$-rays (70–300 MeV) | C.E.N., Saclay (J. Labeyrie), University of Milan (G. Occhialini), Max Planck Institute, Garching (R. Lüst). |

The effort to give flight opportunities to at least some of the experiments of TD2 led to the ESRO IV project, adopted in the spring of 1969 and due for launch in 1972. The spacecraft is similar to ESRO II but the experiments are quite different. The scientific mission is aimed at measuring simultaneously the neutral and ionized components of the atmosphere, together with the low-energy particles producing the ionization. In addition, solar-particle detectors are added to the payload. The experiments are as follows:

| | |
|---|---|
| Positive ion probe | University College, London (R. L. F. Boyd) |
| Mass spectrometry of neutral gases | Bonn University (U. von Zahn) |
| Low-energy particles | Kiruna Geophysical Observatory (B. Hultqvist) |
| Solar particles | Space Research Laboratory, Utrecht (C. de Jager) and Max Planck Institute, Garching (R. Lüst). |

Also adopted in 1969 was a successor to HEOS A-1, the HEOS A-2 project, due for launch in late 1971. The orbit is similar in shape, but polar, where that of HEOS A-1 is essentially equatorial. This means that it will be the first vehicle to penetrate into the area of the neutral point at the border of Earth's magnetic field and that of interplanetary space. The experiments are:

| | |
|---|---|
| Magnetic field measurement | Imperial College, London (H. Elliot) |
| Plasma measurement | University of Rome (G. Pizzella) |
| Solar vlf observation | Danish Space Research Institute (B. Peters) |
| Particle-counter telescope | Space Science Department, ESTEC (E. Page) |
| High-energy electrons | C.E.N., Saclay (J. Labeyrie) and University of Milan (C. Occhialini-Dilworth) |
| Solar wind measurement | Max Planck Institute, Garching (H. Rosenbauer) |
| Micrometeorite detector | Max Planck Institute, Heidelberg (J. Zähringer). |

The early months of 1969 were occupied with defining ESRO's next major projects. A very full investigation of eight possible satellites, with full consultations to avoid scientific duplication with NASA's efforts, led to the adoption of two projects in mid-1969:

COS B, due for launch in 1974 (the payload consisting of a spark chamber to measure galactic $\gamma$-radiation and its energy spectrum) and a geostationary magnetospheric satellite devoted to studies of fields, waves and particles in the magnetosphere at a distance of 6·5 earth radii, due for launch in 1975.

## 5. Sounding rockets

ESRO's activities also include scientific investigations by means of sounding rockets. So far more than 100 rockets have been launched, most of them carrying experiments to study the earth's atmosphere in the region between 50 and 200 km, which is not accessible by satellites or balloons.

The two main rocket ranges from which the ESRO launchings are performed are ESRANGE at Kiruna, Sweden, and the Italian launching site in Sardinia. ESRANGE is particularly suited for investigations of high-latitude phenomena such as aurora borealis or polar-cap absorption events, while Sardinia provides advantages for astronomical investigations.

Significant results obtained so far include:

1. New and more accurate measurements of the neutral composition of the upper atmosphere.

2. Progress in the understanding of the noctilucent cloud phenomenon.

3. Improvement in the understanding of the phenomena which are important for the ionization of the lower atmosphere, by probing the ionosphere during a solar eclipse.

4. Electron and ion density determination of the high-latitude ionosphere during quiet and disturbed conditions.

5. New and more precise measurements of energy spectra of auroral particles down to energies of a few hundred electron volts.

6. Spectrophotometric observations of the aurora borealis in the ultra-violet region.

7. Electric field measurements at medium and high latitudes during quiet and disturbed conditions, employing barium cloud releases.

8. Correlation studies of phenomena observed simultaneously in the solar wind, magnetosphere and high latitude ionosphere, employing data obtained from ESRO I, ESRO II and HEOS satellites related to rocket measurements.

9. Spectrophotometric observations of the night sky in the ultra-violet region.

10. Precise data on the chemical composition of the solar atmosphere by means of X-ray spectrometry.

ESRANGE is not only used for the launching of sounding rockets in ESRO's own programme, but the facilities are also made available for national programmes of member countries. Considerable use of ESRANGE has in fact been made by British, German and Swedish teams.

## 6. Conclusion

It will be seen that ESRO, after rather a long running-in period, is now producing results for European scientists at a significant rate. The facilities, and even more the human resources of the Organisation, now form a well-functioning base for the activities of scientists. Though financial limitations will in the future as in the past necessitate painful selection amongst excellent scientific proposals, European space science is now really able to count on the existence and capability of ESRO.

There are several by-products to this scientific work. As one of ESRO's chief aims, the leadership of science has served, is serving and will serve to advance the technological potential of Europe's industry. This advance is now enabling ESRO to add activities in space applications to its scientific work.

ESRO's international character has many advantageous consequences. The formation of multinational teams in the Organisation is so easy as to be hardly worth mentioning, but to increase the stock of professional personnel experienced in international collaboration is in itself a good thing. A little more difficult is to ensure good collaboration between the industries of different member countries, and here again we have achieved a great deal. Far more difficult is the effective collaboration between ten governments, each with its own problems, crises and elections, each with its own need to justify internally devoting money to external co-operation. Yet here again ESRO can chalk up considerable advances in this most intricate field. Thus we look with great confidence to continuing our work.

A. S. HUDSON

Magnetic Materials Research Group, The Marconi Company Limited, Billericay, Essex

# Molecular engineering in the design of microwave ferrimagnetic garnets

**Abstract.** Recent advances in the study of microwave ferrites and garnets have resulted in a better understanding of the relationship between magnetic properties and the fundamental crystal structure. This article presents a review of the way in which the crystal structure of microwave garnets may be designed in order to realize the many different combinations of magnetic properties required by the designer of non-reciprocal microwave ferrites. The discussion relates to rare-earth and yttrium-containing garnets of the form $Y_{3-a}R_aFe_5O_{12}$ (R = rare earth Sm–Lu) and considers the effect of substitutions of Al, Ga, In, Mn, Ce, Ca and V on magnetization, gyromagnetic resonance, magnetic loss, high-power instability threshold and magnetostriction effects.

## 1. Introduction

The exploitation in the late 1940's of the families of magnetic spinels based upon the mixed oxide formula $M^{2+}Fe_2O_4$ permitted the birth of the now flourishing microwave ferrite device industry. These spinels were the first and simplest of the range of magnetic mixed oxides known as ferrites. In the microwave sphere the early materials included spinels based on magnesium ferrite ($MgFe_2O_4$) and nickel ferrite ($NiFe_2O_4$). The versatility of these two ferrite materials was quickly increased by alterations to the basic compositions, in particular the inclusion of manganese and aluminium in magnesium ferrite (yielding the general formulation $Mg_{1-x}Mn_xFe_{2-y}Al_yO_4$), and the incorporation of aluminium in nickel ferrite to give $NiFe_{2-y}Al_yO_4$.

Although these two ferrites are still of prime importance today for many applications, their place has now been taken to a considerable extent by a second family of ferrimagnetic mixed oxide materials, the magnetic garnets. The discovery of the parent material—yttrium iron garnet ($Y_3Fe_5O_{12}$)—in 1956 (Bertaut and Forrat 1957, Geller and Gilleo 1957) resulted in a fundamental change of philosophy in the design of microwave ferrite devices, and made it possible not only to extend considerably the range and types of devices which may be designed, but also to effect many improvements in their performance and capability.

In parallel with the expansion of the number of different ferrite materials which are available to the device designer, there has been a corresponding advancement in the understanding both of the fundamental physical phenomena which determine the microwave behaviour of ferrites, and of the principles on which device design is based. This advance in understanding now enables the device designer to specify very closely the ferrite properties he requires in order to develop a device for a particular application, and also enables the materials engineer to 'tailor make' the ferrite, designing the material from the basic building block of the unit cell through to the ceramic forming process which will be used to manufacture the final piece parts. It is the purpose of this article to review the current state of the art of 'molecular engineering' of microwave garnets and to show how the molecular formula unit may be designed to realize the desired electromagnetic properties.

## 2. Microwave devices and ferrite properties

Before considering the principles of ferrite material design it is necessary to define the parameters characterizing the ferrite, and to review briefly their importance in relation to the operation of a microwave device.

The devices relevant to this discussion are (i) resonance isolators, (ii) junction circulators, and (iii) phase shifters and phase-shift circulators. These three types of device are passive multiport microwave components which exhibit either (*a*) non-reciprocal attenuation, or (*b*) non-reciprocal phase shift, between adjacent ports. The 3-port junction circulator is shown schematically in figure 1, where it may be seen that the device permits the transmission

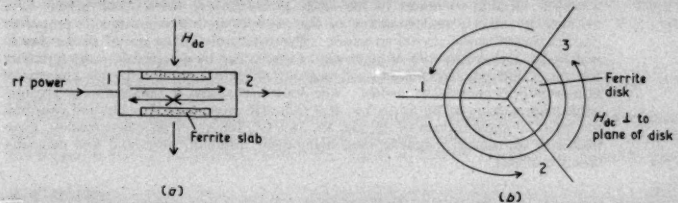



Figure 1. Microwave garnets are used to construct passive non-reciprocal waveguide or stripline components which transmit rf power only in the directions indicated. 2-port devices include the resonance isolator, figure 1(*a*), and one example of a multi-port device is the 3-port junction circulator, figure 1(*b*).

of microwave power with low attenuation from ports 1 to 2, 2 to 3 and 3 to 1, whereas in the reverse direction virtually no power may be transmitted from ports 1 to 3, 3 to 2 or 2 to 1. The isolator is logically a special case of the 3-port circulator with one port omitted; its effect is therefore to transmit power with low attenuation in one direction and to absorb virtually all the power travelling in the opposite direction. The successful design of these devices therefore depends upon achieving two objectives simultaneously: (i) maximum isolation or phase shift in the one direction, and (ii) minimum transmission loss in the reverse direction.

The operation of most devices is dependent in some way or other on the phenomenon of ferromagnetic resonance, whereby energy may be coupled from a circularly polarized rf wave to sustain precession of an unpaired electron about a static applied dc field. This effect is illustrated in figure 2, where the electron is shown precessing with frequency $\omega$ about the internal static dc field ($H_1$) at an angle $\phi$. If the incident rf wave is of the same sense of polarization and the same frequency as the precession, then energy is coupled from the wave to increase $\phi$ and the precession is maintained.

The frequency $\omega$ is called the resonance frequency and depends on the strength of the internal dc field $H_1$ in the ferrite:

$$\omega = \gamma H_1$$

where $\gamma$ is the gyromagnetic ratio.

$H_1$ is in turn a function of the externally applied field $H_A$ and the internal demagnetizing field $H_d$. The internal demagnetizing field arises from the field induced in the sample which tends to oppose the external field, and which is a function of sample shape:

$$H_d = N 4\pi M_s$$

where $4\pi M_s$ is the magnetization and $N$ is the shape demagnetizing factor. $N$ may be resolved into three orthogonal components $N_x$, $N_y$, $N_z$, such that

$$N_x + N_y + N_z = 1$$

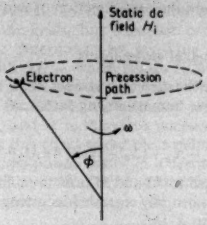



Figure 2

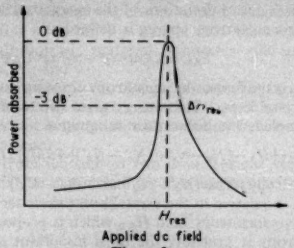



Figure 3

Figure 2. Non-reciprocal power absorption in a microwave garnet is determined by ferromagnetic resonance, whereby power is coupled from the circularly polarized rf wave to maintain precession of a spinning unpaired electron at frequency $\omega$ about a static dc magnetic field $H_i$. This resonance occurs only when the rf polarization is of the same sense as the precession direction.

Figure 3. The shape of the resonance absorption curve as a function of applied magnetic field $H$ for constant frequency $\omega$ is estimated from the width $\Delta H$ of the absorption peak at the half-power (−3 dB) points.

and the internal field in the ferrite is then

$$H_A - N_i 4\pi M_s.$$

The general relation for the resonance frequency is obtained by solution of the equation of motion for the magnetization vector in the finite sample defined by $N_i$, and results in the well-known Kittel equation:

$$\omega = \gamma[\{H_A - (N_z - N_x)\,4\pi M_s\}\{H_A - (N_z - N_x)\,4\pi M_s\}]^{1/2}.$$

This equation defines the resonance frequency for a given sample, magnetization and applied field, and relates to the condition where the direction of circular polarization of the microwave field has the same sense as the precession direction of $M$. Under these circumstances maximum power is absorbed by the ferrite from the microwave field at the frequency $\omega$, whereas when the polarization direction is of opposite sense to that of the precession there is no interaction. A plot of power absorbed as a function of applied field at constant frequency results in the resonance absorption line shown in figure 3; the width of the resonance line ($\Delta H_r$) is of great importance in the design of ferrite non-reciprocal devices since it implies the magnetic losses not only at resonance but also in the immediate vicinity. The parameter representing this width is the '3 dB resonance linewidth', measured at a point where the absorption is half its maximum value (figure 3). Generally it is desirable that $\Delta H_r$ be minimized, implying maximum absorption, and this consideration is particularly valid in the case of the resonance isolator, where it is desirable to obtain a maximum ratio of attenuation in the isolation direction to that in the transmission direction. The theoretical limit for this ratio is given by the relation

$$F = \left(\frac{4\omega}{\gamma \Delta H}\right)^2$$

which takes a maximum value when $\Delta H$ is a minimum.

We have seen that an internal demagnetizing field arises as a function of sample shape to modify the externally applied dc field required for resonance, and there is a second cause which gives rise to an analogous demagnetizing field. This effect is due to magnetocrystalline anisotropy. The anisotropy arises as a result of the variation in magnetic behaviour with direction in the crystal lattice, and may be considered in terms of the energy difference

which arises out of deviations of the magnetization $M$ from the easy direction of magnetization. This anisotropy energy is defined as

$$E_{An} = K_1(\alpha_1^2\alpha_2^2 + \alpha_2^2\alpha_3^2 + \alpha_3^2\alpha_1^2) + \text{higher-order terms}$$

where $K_1$ is the first-order anisotropy constant and $\alpha_i$ are the direction cosines of $M$ relative to the crystal axes. This energy gives rise to an effective demagnetizing factor $N_i'$ which may be included in the resonance equation:

$$\omega = \gamma\{H_A - (N_z - N_x + (N_z' - N_x')\}\, 4\pi M_s]\,[H_A - \{N_z - N_y + (N_z' - N_y')\}\, 4\pi M_s]^{1/2}.$$

The anisotropy terms $(N_z' - N_y')$ and $(N_z' - N_x')$ are related to $K_1$ and $M_s$ and are a function of direction relative to the crystal axes; these terms are normally regarded as giving rise to an effective anisotropy field $H_{An}$ which is proportional to $K_1/M_s$.

Anisotropy is probably the most important single effect determining the behaviour of ferrimagnetic materials, since it not only modifies the resonance condition but also determines the width of the resonance line and leads to the occurrence of high magnetic losses in the unsaturated ferrite at low frequencies. These latter losses, known as 'low-field losses', occur at all frequencies up to $\omega_m$ where

$$\omega_m = \left|\frac{2K_1}{M_s}\right| + 4\pi M_s$$

and it is therefore necessary to minimize $K_1/M_s$ particularly where the material is to be used at low frequency. In considering the width of the resonance line it is necessary to distinguish two contributions, one arising from anisotropy and intrinsic effects, and the other from porosity and other inhomogeneities. These effects will be considered in more detail below.

We have seen how energy is coupled from the rf wave and is dissipated in maintaining the precession angle $\phi$, which increases with the power. This coupling to the uniform precession mode is only valid up to a certain rf power level, and beyond this critical rf field strength power may be coupled from the uniform precession and utilized in the propagation of modes known as 'spin-waves'. This is the so-called 'non-linear' effect and is of fundamental importance where devices are required to handle high rf power levels. If non-linearity sets in, the device becomes excessively lossy, the resonance declines and the device no longer fulfils its function. The rf power level at which this effect sets in is known as the non-linearity threshold $P_{crit}$, and is very much a function of frequency and configuration. The effect of an increase in power through $P_{crit}$ is shown in figure 4, where it is apparent that as $P \to P_{crit}$ the main resonance becomes saturated, and a subsidiary resonance appears at a field below $H_{res}$ for the uniform mode. The threshold power or more usually, the

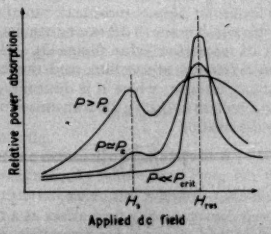



Figure 4. The profile of the resonance curve is strongly dependent upon the power level in the incident rf wave; at a critical power level $P_c$ the main resonance at applied dc field $H_{res}$ degenerates and a subsidiary absorption peak occurs below the main resonance at an applied field $H_s \simeq \omega/2\gamma$. The rf field is perpendicular to $H_A$.

threshold field $h_c$, at which these effects set in, depends upon the relative directions of the rf and dc fields, but whatever configuration is chosen the critical field for non-linearity may be determined by invoking a pseudo-intrinsic material parameter—the spin-wave linewidth $\Delta H_k$—which relates $h_c$, $4\pi M_s$, $\omega$, $H_A$ and $\gamma$.

The importance of non-linear effects lies not only in the change of operating mode of the device, but also in the increased magnetic losses which are encountered when non-linearity sets in, and it is therefore necessary to adopt a material with sufficiently large $\Delta H_k$ to prevent this effect under the conditions envisaged.

The low-power losses in a ferrite arise from the sum of two contributions, one due to magnetic and the other to dielectric loss. These parameters are measured as the imaginary component of the permeability or permittivity respectively:

$$\mu = \mu' + \mu'' \qquad \tan \delta_m = \mu''/\mu'$$
$$\epsilon = \epsilon' + \epsilon'' \qquad \tan \delta_m = \epsilon''/\epsilon'.$$

A detailed knowledge of these loss parameters, together with the static and microwave parameters discussed before, should enable the ferrite material to be sufficiently well characterized for the designer to optimize device performance. For practical purposes it is necessary to know the variation of these parameters with biasing field, frequency and temperature, before real devices may be constructed. Considerations of bandwidth and operating temperature range modify the conclusions so far drawn concerning the optimum values that these parameters should take, but it is not the purpose of this discussion to describe the principles of device design but rather to show how the ferrite material engineer can design ferrites to fulfil almost any required combination of ferrite properties.

## 3. Molecular engineering in the design of microwave garnets

It has been seen in the previous section that in order to arrive at a satisfactory garnet for a particular application it is necessary to control a number of intrinsic properties of the material; these properties include:

| | |
|---|---|
| (i) saturation magnetization | $4\pi M_s$ |
| (ii) Curie temperature | $T_C$ |
| (iii) slope of $4\pi M_s$ against temperature curve | $\mathrm{d}M/\mathrm{d}T$ |
| (iv) resonance linewidth | $\Delta H_r$ |
| (v) spin-wave linewidth | $\Delta H_k$ |
| (vi) intrinsic linewidth | $\Delta H_i$ |
| (vii) anisotropy constant | $K_1$. |

In addition to these parameters, for certain applications it is necessary to control the following:

| | |
|---|---|
| (viii) remanence ratio | $R = B_r/B_s$ |
| (ix) coercive force | $H_c$ |
| (x) magnetostriction coefficients | $\lambda_{100}$, $\lambda_{111}$. |

All these parameters are closely dependent upon the crystal structure of the material, and therefore upon the formulation of the molecule, and also to a lesser extent upon the method of preparation used to manufacture the material in the form in which it is required. Generally speaking the preparation parameters are less important because it is assumed that the optimum preparation process is selected to yield a dense stoichiometric ceramic, and provided that this is the case the final properties depend solely on the design of the basic molecular structure. The only important exception to this rule is the final grain size of the polycrystal which does have very important consequences for some of the parameters. Before consideration is given to the subject of grain size, we will discuss the principles which determine the choice of composition.

3.1. *Control of magnetization*

The material upon which all the subsequent discussion will be based is yttrium iron garnet (YIG)—$Y_3Fe_5O_{12}$. The crystal structure of YIG is cubic (Geller and Gilleo 1957) and belongs to space group number 230 ($Ia3d$). The unit cell contains 8 formula units of type $Y_3Fe_5O_{12}$, and the cations are divided as follows:

$8\times3$ $Y^{3+}$ ions—24 *c*—(dodecahedral) sites
$8\times2$ $Fe^{3+}$ ions—16 *a*—(octahedral) sites
$8\times3$ $Fe^{3+}$ ions—24 *d*—(tetrahedral) sites.

The molecular formula is usually expressed in the following form in order to distinguish the site locations of the ions:

$\{Y_3\}\,[Fe_2]\,(Fe_3)\,O_{12}$ (see figure 5(*a*))
*c*-sites *a*-sites *d*-sites

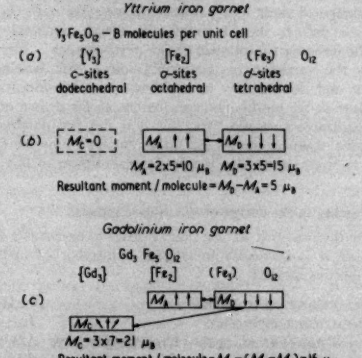



Figure 5. Cations in the garnet crystal structure are arranged on three crystallographic sublattices, figure 5(*a*); in yttrium iron garnet (YIG) only the *a*- and *d*-sublattices contain magnetic ions, figure 5(*b*), whereas in gadolinium iron garnet, magnetic Gd ions are situated on the *c*-sublattice in addition to Fe ions occupying *a*- and *d*-sites, figure 5(*c*).

Each set of equivalent lattice sites represents an individual sublattice, one (the *c*-sublattice) comprising Y ions only, and two (the *a*- and *d*-sublattices) comprising Fe ions only. The magnetic ions align themselves in such a way that the magnetic moments of the ions on any one sublattice are parallel, and the *a*- and *d*-sublattice moments so arising are aligned such that the *a*-sublattice moment is anti-parallel to that of the *d*-sublattice. This situation is illustrated in figure 5(*b*). The overall magnetic moment due to any one sublattice is the sum of the moments of the individual ions on that sublattice:

$$M_A = n_A \times m_A$$

where $n_A$ is the number of magnetic ions in *a*-sites and $m_A$ the individual moment of an *a*-site ion.

The overall magnetization due to the complete crystal is then the resultant of the two sublattice magnetizations:

$$M = M_D - M_A.$$

In the case of YIG, the moment due to one $Fe^{3+}$ ion at 0 K is 5 $\mu_B$ (5 Bohr magnetons), and the overall magnetization of YIG due to both sublattices is therefore

$$M = M_D - M_A (= 15 - 10) = 5\ \mu_B.$$

At any temperature other than 0 K, the average moment $m_A$ of the $Fe^{3+}$ ions will be less than 5 $\mu_B$ due to thermal disorder, but the relation

$$M = n_D m_D - n_A m_A$$

or

$$M = 3m_D - 2m_A$$

is still valid.

Since yttrium is non-magnetic, there is no moment due to the *c*-sublattice, but it is very common to encounter garnets in which the yttrium has been partially or completely replaced by a magnetic rare-earth ion such as gadolinium. The process of replacement of one ion by another is known as substitution, and this is the essence of molecular design. In the case of replacement of Y by Gd, we have the general formula

$$\{Y_{3-a}Gd_a\}\,[Fe_2]\,(Fe_3)\,O_{12} \qquad 0 \leqslant a \leqslant 3$$

which specifies a range of solid solutions between the two end members YIG (when $a=0$ and GdIG (when $a=3$). In this situation there are three magnetic sublattices, whose relative orientations are shown in figure 5(*c*). The principal additional interaction is between *c*- and *d*-sublattices, and the overall magnetization of the molecule is then given by

$$M = |M_C - (M_D - M_A)| \tag{1}$$

where the dodecahedral moment is $M_C = n_C m_C$.

The moment of $Gd^{3+}$ at 0 K is $m_C = 7\ \mu_B$, and consequently when $a=3$,

$$M\{= 3\times 7 - (3\times 5 - 2\times 5)\} = 16\ \mu_B.$$

Since $a$ may take any value between 0 and 3, it is clear that the overall magnetization may be continuously adjusted between the value for $a=0$ (5 $\mu_B$) through zero (at $a=\frac{1}{2}$) to 16 $\mu_B$ at $a=3$ (see figure 6(*a*)).

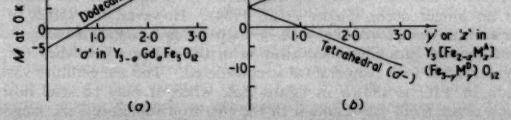



Figure 6. The magnetization $M$ of an individual sublattice in $\{Y_3\}\,[Fe_2]\,(Fe_3)\,O_{12}$ may be altered by (i) substitution of magnetic rare-earth ions for *c*-site Y, figure 6(*a*), or (ii) substitution of non-magnetic ions for *a*- or *d*-site Fe, figure 6(*b*).

There are, however, disadvantages associated with Gd substitution, which will be discussed later, and there are two alternative methods available for controlling magnetization, via substitution in the *a*- and *d*-sublattices. Generally it is not possible to substitute magnetic ions in either the octahedral or the tetrahedral sites, but non-magnetic substitution in these sites is very commonplace. It is therefore possible to postulate the molecular formula

$$\{Y_{3-a}Gd_a\}\,[Fe_{2-z}M_z^A]\,(Fe_{3-y}M_y^D) \qquad O_{12}$$

in which two non-magnetic trivalent ions $M^A$ and $M^D$ have been selected, one which is assumed to prefer octahedral substitution for iron, the other of which is assumed to prefer

tetrahedral substitution. In practice this state of affairs is readily realized by choosing $M^A$ = indium and $M^D$ = gallium.

Considering the situation when $a=0$, we recall that the overall magnetization is

$$M = M_D - M_A = 3 \times m_{te} - 2 \times m_{te}.$$

If Ga is now substituted tetrahedrally, we have

$$M = (3-y)\, m_{te} - 2m_{te}$$

so we see that as tetrahedral substitution increases the *d*-sublattice moment is reduced, thus reducing the overall magnetization, until the point is reached at $y=1$ where the two sublattice magnetizations exactly cancel and $M=0$. Beyond $y=1$, the *d*-sublattice magnetization is smaller than that due to the *a*-sublattice, and, as $y$ is increased beyond 1·0, the overall magnetization rises to reach a maximum value of 10 $\mu_B$ at $y=3$.

Now considering the effect of purely octahedral substitution, we see that as $z$ is increased from zero to its maximum value of 2 atoms/molecule we have

$$M = M_D - M_A\{=3\times 5-(2-z)\times 5\} = 5(1+z)\ \mu_B$$

and the overall magnetization increases linearly to a maximum of 15 $\mu_B$. The two situations of *a*- and *d*-site substitution are illustrated in figure 6(*b*). There is no reason why simultaneous *a*- and *d*-site substitution should not be made, and the resultant magnetization may be obtained by combining the two lines in the figure, giving an infinite number of pairs of $x$, $y$ values to yield any overall magnetization between $-10$ and $+15$ $\mu_B$ at 0 K. Which combinations are selected in practice will depend on other considerations to be discussed later.

The situation where magnetic rare-earth ions are incorporated on the *c*-sublattice in combination with non-magnetic substitution in the *a*- and *d*-sites is of particular interest. The resultant magnetization in this case is obtained from the algebraic sum of the curve in figure 6(*a*) and the resultant of the curves in figure 6(*b*), or alternatively, the value predicted by equation (1). The theoretical maximum magnetization would be obtained when $M_D=0$ and $M_C$ and $M_A$ were a maximum, namely, at $a=3$, $z=0$, $y=3$, where we might expect $M(=3\times 7+2\times 5)=31$ $\mu_B$. This situation is not in fact realized in practice because each sublattice requires a certain amount of magnetic ions on the sublattice with which it interacts in order to maintain its own ferrimagnetic order. Thus, when the magnetization of one sublattice is reduced to zero by high levels of substitution, the ideal ferrimagnetic interactions no longer occur (Geller *et al.* 1964). However, for values of *a*- or *d*-site substitution normally employed (i.e. $y \leqslant 1\cdot 8$, $z \leqslant 0\cdot 7$), equation (1) is valid.

The importance of gadolinium substitution in the *c*-sites arises when the variation of overall magnetization with temperature is considered. The temperature variation of the magnetization in YIG is shown in figure 7(*a*), where it may be seen how the overall magnetization arises from the resultant of the two iron sublattices, the magnetizations of which are identical functions of temperature. When a rare-earth ion is introduced into the *c*-sublattice, the magnetization of the third magnetic sublattice so formed has a completely different temperature characteristic—in general the effective moment of the rare-earth ion falls much more rapidly with temperature than does that of the $Fe^{3+}$ ion. Since both $Fe^{3+}$ sublattices have the same temperature dependence, the situation may be described by plotting the *c*-sublattice moment and the resultant $a-d$ moment together, as shown in figure 7(*b*). In the case of $Gd_3Fe_2(Fe_3)O_{12}$, the *c*-sublattice moment of 21 $\mu_B$ is dominant over the $a-d$ resultant of 5 $\mu_B$ at 0 K, but the rapid decay with temperature of the $Gd^{3+}$ moment means that by some temperature $T_{cp}$ the *c*-moment has fallen to the value of the $a-d$ moment and the resultant overall magnetization is zero. The temperature $T_{cp}$ is known as the compensation point. Beyond the compensation point the $a-d$ moment becomes increasingly dominant as the *c*-moment continues to decay more rapidly. Consequently the overall magnetization rises for a time until the Curie point is approached in the

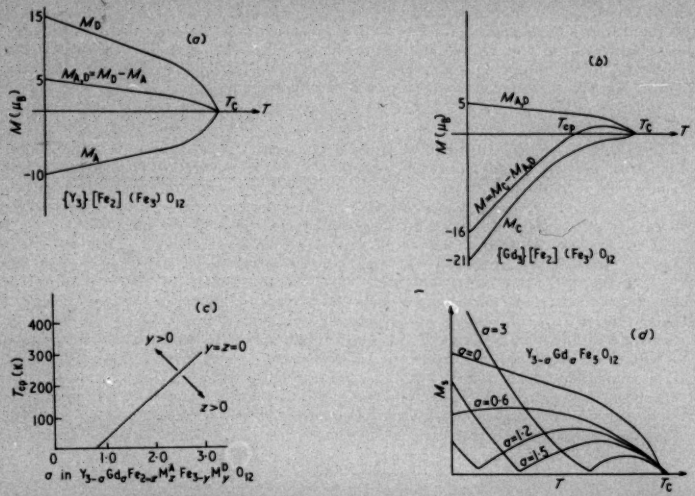



Figure 7. The overall magnetization against temperature characteristic for a garnet is obtained from the resultant of the individual sublattice magnetizations; the two Fe-containing sublattices in YIG decrease at the same rate with temperature, figure 7(*a*), whereas in GdIG the more rapid decrease with temperature of the third Gd sublattice magnetization $M_c$ gives rise to a compensation point $T_{cp}$ below the Curie temperature $T_C$ where the resultant magnetization is also zero. This compensation temperature may be decreased by non-magnetic substitution for *a*-site Fe, or increased by non-magnetic substitution for *d*-site Fe, figure 7(*c*), while partial replacement of *c*-site Gd by Y gives rise to a family of magnetization curves as shown in figure 7(*d*).

usual way. The value of $T_{cp}$ is a function of $a$ in the solid solution series $\{Y_{3-a}Gd_a\}$ [Fe$_2$] (Fe$_3$) O$_{12}$; for $Gd_3Fe_5O_{12}$ (where $a=3$) the compensation temperature is at 290 K. The lowest value of $T_{cp}$ is clearly 0 K and the composition displaying this characteristic is given by the condition $M_C=M_D-M_A=5\ \mu_B$, or $a=\frac{3}{7}$. Variation of $a$ in the range 0·71 to 3·0 thus gives a linear variation of $T_{cp}$ from 0 to 290 K, as shown in figure 7(*c*). The magnetization against temperature characteristics of the members of this family of materials are shown in figure 7(*d*). The significance of Gd substituted YIG is demonstrated by the magnetization characteristic above the compensation point, where $M$ passes through a maximum before falling to zero at the Curie temperature $T_C$. There is therefore a temperature range over which the magnetization is approximately constant and the materials displaying maxima in the range $-100-+100$ °C are of particular interest for device applications.

It is clear from figures 7(*c*) and 7(*d*) that relatively high Gd substitutions are required in YIG to produce materials with $dM/dT=0$ in the vicinity of room temperature. High Gd substitutions are not always desirable because of their adverse effect on resonance properties and therefore non-magnetic substitutions in the *d*-sublattice are used to reduce the $a-d$ moment and so increase the value of $T_{cp}$ for any given Gd substitution. This may be done conveniently (Harrison *et al.* 1962, Maguire and Green 1962) by incorporating $Al^{3+}$ or $Ga^{3+}$ to give the solid solution series:

$$\{Y_{3-a}Gd_a\}\,[Fe_2]\,(Fe_{3-y}M_y)\,O_{12}$$

where M = Al or Ga. Thus putting $y = 0{\cdot}5$ gives $M_{A,\,D} = 2{\cdot}5\ \mu_B$ at 0 K, and $T_{cp} = 0$ K will result from a Gd content of $a = 2{\cdot}5/7 = 0{\cdot}355$ instead of 0·71 when $y = 0$. Similarly for any value of $T_{cp}$ the required amount of Gd is less. The effect of octahedral substitution is naturally the opposite; if In is introduced to give $\{Y_{3-a}Gd_a\}\,[Fe_{2-z}In_z]\,(Fe_3)\,O_{12}$ then $T_{cp}$ is reduced for constant $a$ as $z$ is increased (see figure 7(*c*)).

### 3.2. *Curie temperature*

The slope and nature of the magnetization characteristic about the compensation point (if any) is reflected in the Curie temperature $T_C$ of the material. The magnetization characteristic is controlled by the magnitude and relative rates of decay of sublattice magnetization with temperature, and the Curie temperature is dependent upon the relative concentrations of magnetic ions on the three sublattices in exactly the same way.

The interactions between sublattices discussed above is a result of 'super-exchange' interaction whereby a magnetic ion on one sublattice is linked via an intervening oxygen ion to a magnetic ion on another sublattice. The overall strength of the inter-sublattice interaction so formed therefore depends upon the number of interactions between magnetic ions, and therefore on the number of magnetic ions present.

In pure YIG there are 5 magnetic ions per formula unit, and since the crystal structure shows that each magnetic ion is linked via one intervening oxygen ion to one other magnetic ion, the number of interactions is twice the number of oxygen ions, i.e. 24. The number of interactions per magnetic ion is therefore 24/5. It is found that the ratio of Curie temperatures for substituted and unsubstituted YIG is equal to the ratio of interactions per magnetic ion for the two materials. Thus

$$\frac{T_C(y,z)}{T_C(\mathrm{YIG})} = \frac{n(y,z)}{24/5} \tag{2}$$

where $n(y, z)$ is the number of interactions per magnetic ion in the substituted material.

The value of $n(y, z)$ in a substituted material is not immediately obvious in view of the random distribution of the non-magnetic substituent about the sublattice, but statistical models have been developed (Gilleo 1960) for the evaluation of $n(y, z)$ based upon the probability of any one magnetic ion being linked via oxygen ions to any other. These models result in extremely good correlation between observed and calculated Curie temperatures for $Y_3Fe_{5-y}Al_yO_{12}$, $Y_3Fe_{5-y}Ga_yO_{12}$, and $Y_3Fe_{5-z}In_zO_{12}$.

Subject to certain limitations, similar procedures can be employed to predict the Curie temperature of non-magnetically substituted GdIG and YGdIG (Geller 1960).

Equation (2) indicates that $T_C$ is directly proportional to $n(y, z)$ and therefore non-magnetic substitution will reduce $T_C$ according to the relationship between $n(y, z)$ and $y$ or $z$. At least for substitution less than about 30% of the total magnetic ions on one sublattice, $n(y, z)$ is approximately linear with $y$ or $z$. Thus to a first approximation

$$n(y,z) \propto (3-y) \text{ or } (2-z)$$

but the slope of the line is different for octahedral and tetrahedral substitution, because a given value of $y$ or $z$ (say 0·5 atoms/molecule) represents a different fraction of the total number of magnetic ions on the sublattice (either 25% for octahedral substitution or 16·7% for tetrahedral substitution). The result is that octahedral substitution reduces $T_C$ faster than does tetrahedral substitution. Two examples of this are shown in figure 8, where $T_C$ is shown as a function of Ga substitution ($y$) in $Y_3Fe_{5-y}Ga_yO_{12}$, and In substitution ($z$) in $Y_3Fe_{5-z}In_zO_{12}$. The effect of magnetic substitution in the dodecahedral sublattice is quite different because the number of interactions per magnetic ion is slightly increased by this procedure and the effect is a very slow increase in $T_C$ as the Gd content rises from 0 to 3. This effect is also shown in figure 8.

It should now be clear that it is possible to design a molecular formula based upon the system

$$\{Y_{3-a}Gd_a\}\,[Fe_{2-z}In_z]\,(Fe_{3-y}M_y)\,O_{12} \qquad (M = \mathrm{Al\ or\ Ga})$$

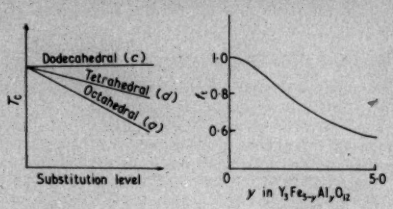



Figure 8 Figure 9

Figure 8. The Curie temperature $T_C$ is also affected by substitution in the three sublattices; magnetic substitution for *c*-site Y produces a slow increase, whereas non-magnetic substitution for *a*-site Fe produces a more rapid decrease in $T_C$ than does substitution for *d*-site Fe.

Figure 9. Al is an example of an ion which does not substitute exclusively in one sublattice; the fraction of Al ions entering tetrahedral *d*-sites as a function of total Al substitution $y$ is expressed by the coefficient $f_t$, figure 9(*a*). This partition of Al between *d*- and *a*-sites causes a slower reduction in magnetization as $y$ is increased than would be expected if all Al ions occupied tetrahedral sites, figure 9(*b*).

which will exhibit any desired magnetization-against-temperature characteristic. For a large number of intermediate characteristics there is therefore a wide choice of different compositions all of which would fulfil most of the desired features. This choice is in fact restricted somewhat by other considerations which are to be discussed in subsequent sections, but before doing this there are some crystallographic considerations to be borne in mind.

### 3.3. *Effect of cation site distribution*

The foregoing analysis depends for its success on an accurate knowledge of the site preferences of the substituent ions. Some substituents are very well behaved in that they substitute exclusively in one lattice site; Gd, In and Ga are examples of such ions. Tetrahedral substitution has been noted earlier as being the most satisfactory method of reducing magnetization, but the element Ga which has so far been selected is inconveniently expensive, and for this reason Al is frequently employed. The disadvantage of Al is that it does not in practice exhibit a total preference for tetrahedral sites, but displays a tendency towards octahedral substitution as $y$ is increased (Geller *et al.* 1964). Thus there is a site distribution of Al ions between octahedral and tetrahedral sites which may be defined by a parameter $f_t$, the fraction of total Al ions which in practice occupy tetrahedral sites. When $y \lesssim 0{\cdot}2$ this parameter is close to unity, but as $y$ exceeds $0{\cdot}2$, $f_t$ falls more and more rapidly as the ions display an increasing tendency towards octahedral substitution. The form of $f_t$ is shown in figure 9(*a*). The practical result of this is that magnetization falls more slowly than would be expected from figure 6(*b*) with increasing $y$, and the Curie temperature falls more rapidly than would be expected from figure 8. The actual variation in practice is shown in figure 9(*b*). However, provided the behaviour of $f_t$ is known for the particular substituent under consideration, the arguments developed above can be quite accurate.

### 3.4. *Control of resonance linewidth*

The importance of the resonance linewidth $\Delta H_r$ as a parameter characterizing the resonance and near-resonance losses in the garnet, was noted above, and it is necessary to consider how this parameter may be maintained at a minimum value by considering how the several contributions to $\Delta H_r$ arise. These contributions include (i) the intrinsic linewidth

of the pure single crystal, (ii) that due to anisotropy and (iii) that due to inhomogeneities, both porosity and secondary phases:

$$\Delta H_r = \Delta H_1 + \Delta H_{An} + \Delta H_p$$

where $p$ is the effective porosity. The minimum value of the overall linewidth may therefore be obtained by minimizing the three contributions separately.

The intrinsic linewidth $\Delta H_1$ is a fundamental parameter of the ferrite system, and for garnets containing no rare-earth ions this contribution is negligible. In the presence of rare earths, however, $\Delta H_1$ may be quite large and is typically in the range 5·7 to 8·4 times the spin-wave linewidth $\Delta H_k$ (Schlömann *et al.* 1960). Thus in the garnet $Y_{1.2}Gd_{1.8}Fe_5O_{12}$, the value of $\Delta H_1$ is 73 Oe, whereas the spin-wave linewidth is about 14 Oe (Dionne 1969 a). Since $\Delta H_k$ is a known function of rare-earth content (see below), the present discussion need only relate to $\Delta H_{An}$ and $\Delta H_p$, the contributions due to anisotropy and porosity.

Several models have been developed for the anisotropy broadening term $\Delta H_{An}$ (Schlömann 1956, 1958, Seiden 1963), and the most appropriate model depends upon the relative magnitudes of the anisotropy field $H_{An} = |2K_1/M_s|$ and the magnetization, and also upon whether an 'independent grain' approach (Schlömann 1958) or a spin-wave model (Schlömann 1956, Seiden 1963) is employed. The independent-grain model, valid where $H_{An} \simeq 4\pi M_s$, gives

$$\Delta H_{An} = \left|\frac{K_1}{M_s}\right|$$

where $K_1$ is the first-order anisotropy constant.
Whereas when $H_{An} \ll 4\pi M_s$ the spin-wave approach gives

$$\Delta H_{An} = 1{\cdot}08\,\frac{J H_{An}^2}{4\pi M_s}$$

where $J = f(H_A, \omega, 4\pi M_s)$. Thus, where the anisotropy field is large, $\Delta H_{An} \propto K_1 M_s$, and where it is small, $\Delta H_{An} \propto K_1^2/M_s^3$. Whichever is the case, the anisotropy contribution to the total linewidth will increase as the magnetization is lowered by non-magnetic substitution, unless the substitution produces a commensurate decrease in $K_1$.

Theoretical models for the origin of anisotropy and for the dependence of $K_1$ upon octahedral and tetrahedral substitution have been investigated by Gerschwind (1961) and Pearson and Annis (1968). Experiment indicates that substitutions of Ga (tetrahedral sites) and Al (tetrahedral and octahedral sites) both lead to a reduction of $K_1$ at room temperature (Dionne 1969 a, table 2), but the effect is more marked in the case of Al, to the extent that $K_1/M_s$ decreases with this latter substitution but increases with the former. Exclusive octahedral substitution (for example indium) leads to the most effective reduction of $K_1$, not only in YIG (Pearson and Annis 1968) but also in the related system $Y_{3-2x}Ca_{2x}Fe_{5-z-z}V_xIn_zO_{12}$ (Van Hook *et al.* 1968), and the latter reference reports the observation of $K_1 \simeq 0$ at room temperature at $x = 0{\cdot}8$, $z = 0{\cdot}5$. For $c$-site substitution (Gd), $K_1$ remains approximately independent of the substitution level (Dionne 1969 a). The observed dependence of $M_s$, $K_1$, and $K_1/M_s$ on $c$-, $a$- and $d$-site substitution of the ions discussed above is summarized in figure 10($a$), and the variation of observed linewidth is shown in figure 10($b$).

The third contribution to the polycrystalline linewidth is $\Delta H_p$, due to air porosity and magnetic inhomogeneities. This contribution, according to the independent grain model, is given by (Dionne 1969 a)

$$\Delta H_p = 1{\cdot}5 \times 4\pi M_s p(1-p)$$

where $p$ is the effective porosity. $\Delta H_p$ is therefore a function of $p$, and the contribution may be eliminated by putting $p = 0$. This may be achieved only by using processing conditions which are optimized to yield 100% dense, homogeneous stoichiometric bodies (Hudson 1969 a). For any given garnet composition the contribution due to porosity can be

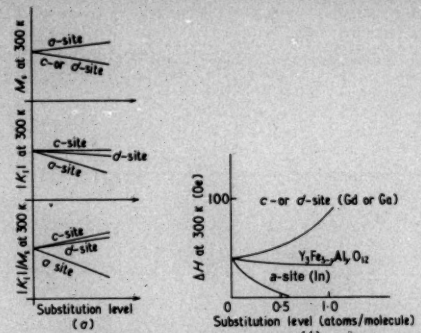



Figure 10. The resonance linewidth, $\Delta H$ (see figure 3) is a function of magnetocrystalline anisotropy $K_1$ and magnetization $M_s$. *c*-, *a*- and *d*-site substitution affects $K_1$ and $M_s$ in different ways, combining to increase $K_1/M_s$ for exclusive *c*- or *d*-site substitution, and to decrease it for exclusive *a*-site substitution, figure 10(*a*). Substituents such as Al (which do not substitute exclusively on one sublattice) produce, in practice, an intermediate effect on the observed linewidth, figure 10(*b*).

calculated exactly, e.g. for a commonly used magnetization of 1000 gauss, $\Delta H_p$ is approximately 15 Oe/per cent porosity. Therefore if it is known that the ceramic production process being employed to manufacture the material will guarantee densities in excess of, say, 99% theoretical x-ray density, a maximum $\Delta H_p$ of 15 Oe is implied. There is, however, the problem of magnetic inhomogeneity, which is not directly measurable from density considerations; this component is usually less or comparable with that due to air porosity, resulting in a conservative estimate of say 20–25 Oe for the overall porosity contribution.

### 3.5. *Control of spin-wave linewidth*

Spin-wave linewidth $\Delta H_k$ determines the critical rf field strength at which power begins to be coupled into the propagation of spin-waves rather than into maintaining the uniform precession mode. $\Delta H_k$ may conveniently be determined in terms of the critical rf field strength $h_c$ under conditions where the applied dc biasing field is parallel to the rf magnetic field vector. This is the 'parallel pump' instability threshold where

$$\Delta H_k = \frac{\gamma 4\pi M_s h_c}{\omega}.$$

$h_c$ is observed experimentally by observing the effect of increasing rf power on the shape of a microwave pulse transmitted through a cavity containing the ferrite sample, to which a field $H_s$, lower than that required for main resonance, is applied. This field is given by

$$H_s = \frac{\omega}{2\gamma}.$$

There are two methods available for control of $\Delta H_k$: (i) introduction of 'fast relaxing' rare-earth ions into the *c*-sublattice (Schlömann 1965), and (ii) reduction of grain size (Whicker and Jones 1966). All rare-earth ions are effective in increasing $\Delta H_k$, but the most effective are holmium, terbium and dysprosium. Holmium and terbium are so powerful

that only very small quantities are required. The relative effects of these ions are shown compared with gadolinium in figure 11(*a*) as a function of substitution level *b* in the materials

$$\{Y_{3-b}R_b\}\,[Fe_2]\,(Fe_3)\,O_{12} \qquad R = Ho, Dy, Gd.$$

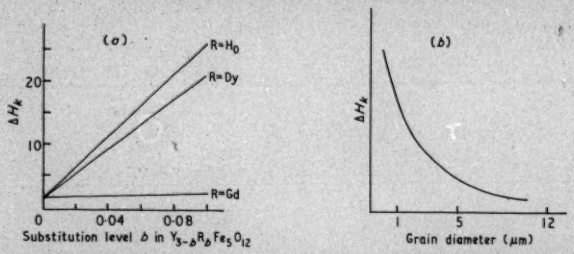



Figure 11. The critical rf power level $P_c$ (see figure 4), at which a subsidiary absorption peak occurs in the resonance absorption curve, is proportional to the square of the spin-wave linewidth $\Delta H_k$; $\Delta H_k$ may be increased by small amounts of rare-earth substitution with varying degrees of effect, figure 11(*a*), or by reduction in mean grain size, figure 11(*b*).

The greatly enhanced power of Dy and Ho is very clear from the figure; since $\Delta H_k$ is directly proportional to $h_c$, and power $P_c$ is proportional to $h_c^2$ it follows that a doubling of $\Delta H_k$ implies a fourfold increase in the rf power which can be tolerated by the ferrite material without encountering non-linearity. In the case of Dy or Ho substitution a concentration of approximately 0·02 atoms/molecule is required to double $\Delta H_k$.

Reduction of grain size has a very similar effect on $\Delta H_k$; in conventional garnets the mean grain diameter is of the order 8–12 $\mu$m and this is the grain size on which the results in figure 11(*a*) are based. For grain diameters greater than 5 $\mu$m, $\Delta H_k$ is independent of the diameter, but below the 5 $\mu$m level dramatic increases in $\Delta H_k$ are observed. This effect is illustrated in figure 11(*b*). It is therefore possible by reduction of grain size to attain $\Delta H_k$ values comparable with those obtained by rare-earth doping, but in practice it is not easy to produce dense ceramics with grain diameters less than 5 $\mu$m, and the rare-earth approach is more frequently used.

Increases in the value of $\Delta H_k$ are always accompanied by an increase in magnetic loss; such increases imply a corresponding increase in the intrinsic linewidth $\Delta H_i$, and it has been found that for a given type of composition the intrinsic and spin-wave linewidths are related (Schlömann *et al.* 1960):

$$\Delta H_i = k\,\Delta H_k$$

and it has already been remarked that for $Y_{3-a}Gd_aFe_5O_{12}$ the value of $k$ is about six. Although $k$ varies slightly from one material to another, it is generally observed that this equation successfully predicts the behaviour of garnets subjected to doping by fast-relaxing ions. This relation is also useful for relating loss and power handling characteristics of materials of the same composition but varying grain size. In this situation

$$\Delta H_i = k\,\Delta H_k = \text{function}\,(1/d)$$

(where $d$ is the mean grain diameter).

A knowledge of the behaviour of $\Delta H_k$ and $\Delta H_i$ as a function of doping level and microstructure consequently enables the precise tailoring of material properties in respect of both power handling capability and the intrinsic contribution to the resonance linewidth.

3.6. *Control of parameters at remanence*

The operation of the majority of microwave devices which work at remanence depends upon the amount of differential phase shift which is obtained from the ferrite in the remanent magnetized state. This phase shift $\Delta\phi$ is dependent upon the two permeabilities $\mu^+$ and $\mu^-$ obtained by switching the magnetization into the two opposing remanent states. Thus

$$\Delta\phi = \mu^+ - \mu^- = f(4\pi M).$$

The value of phase shift depends upon the remanent magnetization which is a property of the hysteresis loop shown in figure 12. The parameters of interest are then the remanence ratio $R = B_r/B_s$, and the coercive force $H_c$. The remanence ratio $R$ is in turn dependent upon the mechanical stress applied to the crystal via the phenomenon of magnetostriction, and the successful operation of the device depends upon $R$ being constant in spite of variation

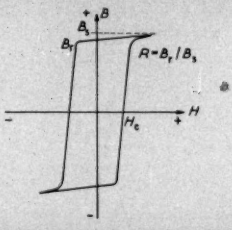



Figure 12. The design of latched microwave devices which operate with the garnet at remanence depends upon the remanent induction $B_r$ and the coercive force $H_c$ which are obtained from a study of the hysteresis loop. The remanence ratio $R = B_r/B_s$ should be high to maximize the differential phase shift obtained by switching the garnet positive or negative, and $H_c$ should be low to minimize the energy required to obtain such switching.

in stress applied to the material. It is possible to determine the relationship between $R$ and the magnetostriction constants $\lambda_{100}$ and $\lambda_{111}$ by considering the changes in anisotropy energy resulting from an applied stress parallel to any direction in the crystal (Dionne 1969); the relationship depends upon the magnitude and sign of $K_1$. For most garnets $K_1$ is negative, and the effect of a compressive stress $\sigma_c$ (applied parallel to the principal crystallographic directions $\langle 100\rangle$ and $\langle 111\rangle$) on the remanence ratio $R$ is given by

$$R^2 = \sigma_c \frac{\lambda_{100}}{K_1} + \tfrac{1}{3} \quad \text{for } \sigma_c \parallel \langle 100\rangle$$

$$R^2 = 1 \qquad \text{for } \sigma_c \parallel \langle 111\rangle$$

and for the intermediate case ($\sigma_c \parallel \langle 110\rangle$)

$$R^2 = \frac{(\lambda_{100} + \lambda_{111})\,\sigma_c}{2K_1} + \tfrac{5}{6}.$$

These relations hold true for stress applied to a single grain, and the behaviour of a polycrystal has to be inferred from the result obtained for the intermediate $\langle 110\rangle$ direction.

The behaviour of $R$ with $\sigma_c$ may be generalized as

$$R^2 = R_0^2 + \Delta R^2$$

where the unstressed remanence ratio $R_0$ takes values limited by $1/\sqrt{3}$ in the $\langle 100\rangle$ direction, through $(2/3)^{1/2}$ in the $\langle 110\rangle$ direction, to the maximum value of 1 for the easy direction $\langle 111\rangle$. The change in $R$ produced by stress is determined in magnitude and sign by the

term $\lambda\sigma_c/K_1$, so that in the case of most garnets (where $\lambda$ and $K_1$ are negative), a compressive stress produces increased remanent magnetization and therefore increased phase shift $\Delta\phi$.

One practical case of interest arises in the construction of a remanent phase shifter utilizing toroidal geometry (Whicker and Jones 1966 a, b, Pippin 1968). In this situation the dissipation of heat (arising from the magnetic and dielectric loss of the ferrite) gives rise to a radial temperature gradient which in turn results in a tensile stress parallel to the magnetization. Since $\sigma_c$ is negative, increased stress will decrease $R$ if $\lambda/K_1$ is positive and increase $R$ if $\lambda/K_1$ is negative. For most applications it is desirable that $\Delta R=0$, and this can only be so if $\sigma_c=0$ or $\lambda=0$. At present, two methods of achieving $\lambda=0$ are known, both of which involve doping, and both of which may reduce either $\lambda_{100}$ or $\lambda_{111}$ to zero.

This reduction in $\lambda$ may be obtained either (i) by *c*-site substitution or (ii) *a*-site substitution. The first case involves the use of cerium (West and Blankenship 1968), giving a molecule of type $Y_{3-c}Ce_cFe_5O_{12}$ and it has been shown that putting $c=0{\cdot}075$ virtually removes the stress dependence of $R$. The second case (Dionne and Paul 1969) utilizes $Mn^{3+}$ as a substituent for octahedral $Fe^{3+}$:

$$\{Y_3\}\,[Fe_{2-z}Mn_z]\,(Fe_3)\,O_{12}$$

which results in the change of sign of both $\lambda_{100}$ and $\lambda_{111}$ as $z$ approaches $0{\cdot}1$ atoms/molecule. It is therefore possible to choose a level of manganese doping which will reduce either coefficient to zero, or alternatively arrange that $\lambda_{100}+\lambda_{111}=0$ as required for the elimination of magnetostrictive effects when $M$ is parallel to $\langle 110\rangle$.

## 4. Conclusions

The preceding discussion has attempted to show how it is possible to adjust almost any intrinsic parameter of a microwave garnet by suitable design of the fundamental molecule. The completely generalized formulation for such a material then becomes

$$\{Y_{3-a-b-c}Gd_aR_bCe_c\}\,[Fe_{2-z-d}In_zMn_d]\,(Fe_{3-y}M_y)\,O_{12}$$

where R = Dy, Ho or Tb and M = Al or Ga.

In this system it is possible to control (i) magnetization and magnetization against temperature characteristic, (ii) resonance properties, (iii) high-power properties, (iv) remanence properties, by suitable adjustment of the molecular composition parameters $a$, $b$, $c$, $d$, $y$, $z$. The effects of these substitutions are summarized in table 1. It is clear from the table that most of the measures which may be taken to optimize the several properties involve a degree of conflict between competing requirements. For example, the anisotropy field $H_{An}$, and consequently the anisotropy contribution to the resonance linewidth, can only be minimized at the expense of a reduction in Curie temperature and an increase in the slope of the magnetization against temperature relation. This implies that a device utilizing a garnet with very low linewidth in order to obtain maximum bandwidth will not in general be capable of operating over a large temperature range, and vice versa. Similarly a high-power device which requires a garnet with a large spin-wave linewidth is immediately burdened with a higher intrinsic linewidth and therefore a higher insertion loss. For these and similar reasons the design of a material which presents an optimum compromise between the various conflicting requirements is a matter for some conjecture on the part of the materials designer, and the successful employment of the considerations presented above are his 'raison d'être'.

In conclusion it is worth mentioning a comparatively recent development in microwave garnets which is ultimately likely to result in the emergence of even better materials. These new garnets are a development of substituted YIG, but utilizing the simultaneous substitution of divalent calcium for yttrium and pentavalent vanadium for iron:

$$\{Y_{3-2x}^{3+}Ca_{2x}^{2+}\}\,[Fe_2^{3+}]\,(Fe_{3-x}^{3+}V_x^{5+})\,O_{12}\text{–YCaVIG}$$

(Hodges *et al.* 1966). In this system the parameter $x$ may take any value from 0 to $1{\cdot}5$,

**Table 1. Summary of garnet properties and means of control**

| Property of material | | Requirement | Determines | Means of control | Side effects |
|---|---|---|---|---|---|
| Magnetization | $4\pi M_s$ | (*a*) Increase<br>(*b*) Decrease | Frequency, bandwidth, loss, phase shift | (*a*) *a*-site substitution—In<br>(*b*) *d*-site substitution—Ga or Al | Reduced $T_C$<br>Increased $dM/dT$ |
| Magnetization against temperature | $dM/dT$ | Low or zero at operating temperature | Operating temperature range | *c*-site substitution—Gd | Increased $H_A$, $\Delta H_A$, $\Delta H_i$, $\Delta H_g$ |
| Resonance linewidth | $\Delta H =$ | Minimum | Insertion loss, bandwidth, phase shift | (*a*) Minimize rare-earth content | Increased $dM/dT$ |
| | $\Delta H_i +$ | | | (*b*) *a*-site substitution—In | Reduced $T_C$ |
| | $\Delta H_A +$ | ,, | ,, | (*a*) *a*-site substitution—In<br>(*b*) *d*-site substitution—Al | Increased $dM/dT$<br>Reduced $T_C$ |
| | $\Delta H_p$ | ,, | ,, | Eliminate porosity and extraneous phases | — |
| Spin-wave linewidth | $\Delta H_g$ | Adjust to high-power requirements | Maximum power level | (*a*) *c*-site substitution—Dy, Ho, Tb<br>(*b*) Reduce grain size | Increased $\Delta H_i$ |
| Magnetostriction constant | $\lambda_{100}$<br>$\lambda_{111}$ | Minimize | Stress dependence of phase shift | (*a*) *c*-site substitution—Ce<br>(*b*) *a*-site substitution—Mn | Increased $\Delta H_i$<br>— |
| Dielectric properties | $\epsilon$<br>$\tan \delta_\epsilon$ | —<br>Minimize | —<br>Insertion loss | Dependent upon substitution<br>Ensure precise stoichiometry | —<br>— |

thus yielding a new family of mixed yttrium–calcium–vanadium garnets. The end member (with $x=1{\cdot}5$) is calcium–vanadium–iron–garnet:

$$\{Ca_3\}\,[Fe_2]\,(Fe_{1{\cdot}5}V_{1{\cdot}5})\,O_{12}\text{–CaVIG}$$

which is particularly noteworthy for the (as yet unexplained) fact that its Curie temperature is 220 °C, some 100–150 °C higher than is the case with other *d*-site substitutions for iron at the 1·5 atoms/molecule level. A related material to CaVIG is obtained by substituting bismuth for calcium:

$$\{Bi_{3-2x}Ca_{2x}\}\,[Fe_2]\,(Fe_{3-x}V_x)\,O_{12}\text{–BiCaVIG}.$$

The two types of material YCaVIG and BiCaVIG thus provide two ranges of mixed garnets in which the magnetization may be varied continuously between $+5\,\mu_B$ and $-2{\cdot}5\,\mu_B$ (at 0 K) by adjustment of the parameter $x$. These materials all exhibit higher Curie temperatures than may be obtained with other non-magnetic substitutions, and the resonance properties are comparable with the more conventional materials. There has been considerable study of these materials, particularly with regard to the effects of *c*-, *a*- and *d*-site substitution (Wilson *et al.* 1967, Hudson 1969 b, Van Hook *et al.* 1968), and it is likely that materials based on YCaVIG and possibly BiCaVIG will ultimately replace some of the more expensive conventional substituted garnets.

**Acknowledgments**

The author wishes to thank the Director of Research, The Marconi Company Limited, for permission to publish this article, and to acknowledge the work of his colleagues of the Magnetic Materials Research Group in the study of microwave materials.

**References**

Bertaut, F., and Forrat, F., 1957, *C.R. Acad. Sci., Paris*, **244**, 96.
Dionne, G. F., 1969 a, *J. appl. Phys.*, **40**, 1839.
—— 1969 b, *IEEE Trans. Magnetics*, MAG-5, **3**, 596.
Dionne, G. F., and Paul, P. J., 1969, *Mat. Res. Bull.*, **4**, 171.
Geller, S., 1960, *J. Phys. Chem. Solids*, **16**, 21.
Geller, S., and Gilleo, M. A., 1957, *Acta Cryst.*, **10**, 239.
Geller, S., Williams, H. J., Espinosa, G. P., and Sherwood, R. C., 1964, *Bell Syst. Tech. J.*, **43**, 565.
Gilleo, M. A., 1960, *J. Phys. Chem. Solids*, 13, 33,
Gerschwind, S., 1961, *Phys. Rev.*, **121**, 363.
Harrison, G. R., Hodges, L. R., Jr., 1962, *J. appl. Phys.*, **33**, 1375.
Hodges, L. R., Rodrigue, G. P., Harrison, G. R., and Sanders, A. D., 1966, *J. appl. Phys.*, **37**, 1085.
Hudson, A. S., 1969 a, *Marconi Rev.*, **32**, 173, 129.
—— 1969 b, *IEEE Trans. Magnetics*, MAG-5, 3 610.
Maguire, E. A., and Green, J. J., 1962, *J. appl. Phys.*, **33**, 1373.
Pearson, R. F., and Annis, A. D., 1968, *J. appl. Phys.*, **39**, 1338.
Pippin, J. E., 1968, *Microwave J.*, 29–45.
Schlömann, E., 1956, *Am. Inst. Elect. Engrs spec. Publ.* T-91, 600.
—— 1958, *J. Phys. Chem. Solids*, 6, 242.
—— 1965, *IEEE Trans. Magnetics* MAG-1, 3, 168.
Schlömann, E., Green, J. J., and Milano, U., 1960, *J. appl. Phys.*, **31**, 386S.
Seiden, P. E., 1963, *J. appl. Phys.*, **34**, 1696.
Van Hook, H. J., Green, J. J., Euler, F., and Czerlinsky, E. R., 1968, *J. appl. Phys.*, **39**, 730.
West, R. G., and Blankenship, A. C., 1968, *Intermag. Proc.*, p. 17.8.
Whicker, L. R., and Jones, R. R., 1966 a, *Microwaves*, 31.
—— 1966 b, *Microwaves*, 43.
Wilson, W. R., Hodges, L. R., Rodrigue, G. P., and Harrison, G. R., 1967, *J. appl. Phys.*, **38**, 1405.

J. E. THOMPSON
Department of Electrical Engineering, University Institute of Technology, Cardiff

# Recent developments in conventional soft magnetic materials

**Abstract.** Recent developments in conventional soft magnetic materials, with particular reference to their applications, are discussed under three main headings: grain-oriented silicon-iron, high permeability materials and thin ferromagnetic films. Ferrites have been excluded since they will be the subject of a later article.

## I. Grain-oriented silicon-iron

### 1. Introduction

The most important commercial soft magnetic material is grain-oriented silicon-iron, which is used in large quantities in the electrical engineering industry. In the early days of the electrical manufacturing industry ordinary sheet steel was in common use, but the magnetic quality of this material was very variable. The development of the modern grain-oriented silicon-iron started in 1900 when Hadfield and his co-workers (Barrett *et al.* 1900) discovered that the addition of silicon, as an alloying element, stabilized and improved the performance of core steel. Further improvements were obtained with better manufacturing techniques; however, subsequent improvements depended on the production of a grain-orientation in the material, which enabled the anisotropic properties to be utilized in a beneficial way.

For the benefit of those who are not familiar with grain-oriented silicon-iron as a material, at this stage it is worth while discussing briefly its salient properties and elaborating on these in later sections.

In the original type of commercial material, made by the process developed by Goss (1934), the degree of preferred orientation was about two-fifths owing to primary recrystallization. The power loss at 50 Hz and 15 kG was about 1·0 W $lb^{-1}$ for 0·012 in thick laminations. Modern 'Goss'-oriented material originated some time later (1940), when it was found possible to increase the degree of preferred orientation by secondary recrystallization, which reduced the loss to about 0·55 W $lb^{-1}$. Subsequent general improvements in purity and manufacturing techniques have reduced this figure to between 0·4 and 0·5 W $lb^{-1}$. To the electrical engineer, the power loss is probably the most important single property of silicon-iron, and much work has therefore been devoted in recent years to reducing this value. This work has explored various avenues; attempts have been made to increase the degree of preferred orientation, to find other alloying elements that could replace silicon either wholly or partially and give better properties. Perhaps most important of all has been the realization that some metallurgical and physical properties are related, and that the desirable properties can be modified (more often than not in a detrimental manner) by the way in which the material is used in electrical machinery. Consequently, most of the recent improvements that have occurred have evolved from the result of interdisciplinary investigations by physicists, metallurgists and electrical engineers.

### 2. Power loss

#### 2.1. *General*

The energy loss in the material at 50 Hz is of prime importance; it is the grading criterion on which the material is sold. Thus, because of its commercial significance, much work

has been carried out in an attempt to understand fully the nature of the loss and to reduce its magnitude.

One important approach has been based on the fact that the power loss at high frequency is always greater than the loss measured in slowly varying fields, which is termed the static hysteresis loss. In rapidly varying fields, eddy currents are set up in the material which result in an additional loss. Unfortunately, the eddy current loss calculated in terms of classical theory does not agree with the experimental value, obtained as the difference between the power loss and the static hysteresis loss. The excess loss is called the 'anomalous' or discrepancy loss and at 50 Hz in silicon-iron it is roughly equal to the classical eddy current loss.

Attempts to explain the anomalous loss have fallen into two categories. The first involves making corrections to the classical equation for eddy current loss. This formula was derived on the assumption of contant permeability, which means that if the applied field is sinusoidal then the eddy current and induction will also be sinusoidal. The relationship between $B$ and $H$ is far from linear, so that if one of these quantities is linear the other must have a distorted waveform. Measurements with sinusoidal $B$ show that an anomalous loss is still present. The effects of flux waveform distortion have been studied in detail by Brailsford (1948). While he showed that flux waveform distortion could account for some of the anomalous loss, from a more recent study with a refined model (Brailsford and Burgess 1961) he concluded that the anomalous loss is not due to the distortion of the internal flux waveform. It would seem that no satisfactory explanation of anomalous loss can be achieved by modifications to the classical theory of eddy current losses.

A second approach has been to calculate the eddy current loss associated with moving domain walls. Pry and Bean (1958) have done this for a system involving a number of parallel 180° walls. The final expression for the loss is complicated and a parameter based on the ratio of domain spacing to specimen thickness is involved. An alternative approach has been developed by Williams *et al.* (1950), based on the damping of domain walls, but Lee (1958) has shown that this treatment leads to the same results as those of Pry and Bean. Leak and Duckworth (1963) concluded that the expression derived by Pry and Bean explained fully the observed eddy current loss in 46 grade grain-oriented 3% silicon-iron for strip thicknesses greater than about 0·010 in; however, in thinner strips, the anomalous loss was higher than expected. They believed that this arose because the effective domain size under dynamic magnetization was greater than that which was observed statically using colloid techniques.

In a polycrystalline sample, it is possible that not all the walls are mobile under alternating conditions: some may be held back by various pinning mechanisms, such as inclusions; indeed, all walls may not move equally quickly. These aspects have been considered recently by Boon and Robey (1967) and Boon (1967), who investigated domain-wall motion in a number of adjacent grains in a sample of grain-oriented silicon-iron. These authors concluded that the anomalous loss could be accounted for, in a satisfactory manner, by the variations in wall mobility. The variation of the anomaly factor with frequency was explained in terms of domain-wall bowing. Unfortunately, the measurements were limited to 6 kG since, above that flux density, the domain-wall motion was non-repetitive.

On domain models, eddy-current shielding effects are greater than expected on the classical model, so that the wall moves further on the surface and is retarded in the centre, progressively so with increasing frequency; hence the origin of the term 'domain-wall bowing'.

Overshott *et al.* (1968) considered that the simplest starting point in any attempt to explain the complex nature of the anomalous loss was to make measurements of domain-wall movements, the flux density and power loss (all as a function of frequency) on an individual grain of the material, since this is the basic unit. The way in which the individual grains integrate in a co-operative manner, to give the average behaviour of the material, then becomes a further problem to be tackled. Measurements of the position of parallel-bar domain walls, at various points on the magnetizing cycle over the frequency range 25–100 Hz, were made using the Kerr magneto-optic technique. Within a number of

well aligned grains of good quality material, the wall-movements were uniform and repetitive. To measure the flux density, two small holes were drilled on opposite sides of the grain and a single-turn search coil was wound around the test region (Brailsford and Mazza 1962). This enabled the true or volume degree of saturation to be obtained by dividing the flux density, measured by the coil, by the saturation flux density (20 kG). The degree of magnetic saturation, as indicated by the domain surface, can be written as $(a-b)/(a+b)$, where $a$ is the combined width of domains magnetized one way, and $b$ that of those magnetized the other way. The walls disappeared from the surface before complete saturation in the volume of the material was reached. The effect can be seen also in figure 1, where the degree of surface magnetization, $(a-b)/(a+b)$, is plotted against the volume magnetization for various frequencies. The amount of domain-wall bowing that exists is surprising.

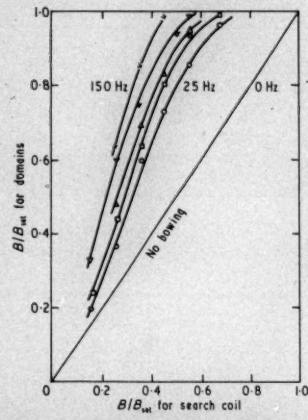



Figure 1. $B/B_{sat}$ as obtained from the domain positions and the search coil, for an individual grain of silicon-iron. ○ 25 Hz; □ 40 Hz; △ 60 Hz; ▽ 100 Hz; × 150 Hz.

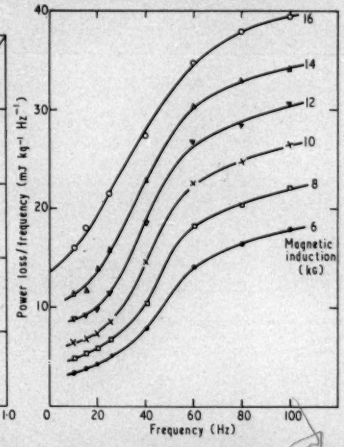



Figure 2. Power loss/frequency against frequency at various magnetic inductions, for the same grain as used in the domain measurements.

The power loss in the grain was measured using a modified form of an apparatus described by Boon and Thompson (1965). The experimental results are shown in figure 2. In general, such experimental results on bulk material are considered to give a linear relationship, but it is worth noting that Brailsford and Fogg (1966) have reported that the loss at lower frequencies is less than might be expected from extrapolation, even for this case, giving a pronounced curvature at the lower frequency end of the curves.

### 2.2. *The variation of power loss with stress and temperature*

A knowledge of the variation of power loss with stress and temperature is important in a number of applications. For example, compressive stresses along the plane of the laminations arise in a large transformer (several hundred MV A) owing to the flattening of wavy laminations (Wilkins and Thompson 1962). These stresses may only be of the order of a few hundred points per square inch, and the material may be very good in terms of flatness by

general metallurgical standards, but such stress levels are sufficient to appreciably alter the loss. Temperature gradients can generate similar stresses in power transformers if the cooling across and through the limbs is insufficient to remove adequately heat generated in the core. Here again the relevant factor is the stress set up in the plane of the lamination and along the direction of flux.

Brailsford and Abu-Eid (1963) have carried out a detailed investigation of the variation of hysteresis loss, permeability, remanence and coercive force of grain-oriented silicon-iron with tensile stresses up to 15 000 lb $in^{-2}$. Measurements, of importance to the design engineer, of the variation in power loss and other magnetic properties with applied compressive and tensile stresses of up to 500 lb $in^{-2}$, and at various temperatures in the range 20–400°C, have been reported by Brown, Holt and Thompson (1965).

The power loss increases by a few per cent for every 100 lb $in^{-2}$ applied compressive stress, and falls at a somewhat lower rate under tensile stress. The power loss falls with increasing temperature, but a temperature difference between the centre and edge of a transformer core of about 150 degC can generate a compressive stress of some 400 lb $in^{-2}$. The decrease in loss owing to a higher temperature can be offset by the compressive stress generated by the temperature gradient.

### 2.3. *Rotational power loss*†

Most studies of power loss in magnetic materials are based on alternating flux measurements, where the direction of the flux remains constant and its magnitude varies cyclically. A different loss process takes place, however, when the magnitude of the flux remains

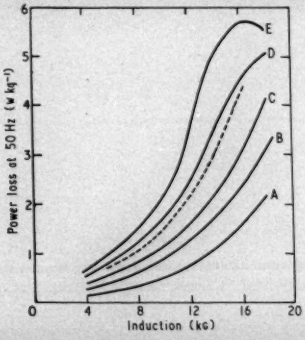



Figure 3. Rotational and alternating losses in 0·013 in 'Goss'-oriented 3% silicon-iron. - - - power loss in transverse direction. Ratio of transverse to rolling-direction flux density shown as percentage: A 0%; B 33%; C 50%; D 75%; E 100%.

constant and its direction varies cyclically, giving a loss known as 'rotational hysteresis loss'. In practice, both alternating and rotational processes occur together in varying amounts in most electrical apparatus, but rotational power loss is particularly important at the T-joints of three-phase transformers and behind the teeth of electrical machines.

Boon and Thompson have carried out a study of alternating and rotational power loss at 50 Hz in 3% silicon-iron sheets.

† For simplicity, the terms 'alternating power loss' and 'rotational power loss' will be used for the losses occurring at 50 Hz under alternating and rotating flux conditions, respectively. The terms 'alternating hysteresis loss' and 'rotational hysteresis loss' will be used for the corresponding losses under slowly varying flux conditions (i.e. essentially static).

The variation of power loss with induction is given in figure 3 for various ratios of the transverse-direction flux density to the rolling-direction flux density for 0·013 in 'Goss'-oriented 3% silicon-iron. It should be noted that the rotational power loss is considerably greater than the alternating power loss. An interesting point that also emerged was that the main advantage of 'four-square' silicon-iron (see §4.2) may be in applications where a large rotational flux is present.

## 3. Magnetostriction and transformer noise

### 3.1. *Introduction*

Although magnetostriction is of considerable interest to the physicist, to the transformer engineer it is an annoying factor giving rise to transformer noise. Transformer noise can be generated in two ways: by the movement of sheets or transformer limbs owing to electromagnetic forces, and by the magnetostrictive changes of the silicon-iron core. The latter predominates in a well-designed core. Although a satisfactory reduction of noise can be achieved by enclosing the transformer in a noise-reducing enclosure, it is worth while to consider how the noise might be reduced at the source.

### 3.2. *The measurement of magnetostriction*

A number of methods have been used to measure the magnetostriction of laminations. On large transformers, vibration transducers are comparatively simple and convenient to use, but for laboratory work, where Epstein samples (30·5 cm × 3 cm) of convenient size are usually available, a technique employing ceramic displacement transducers as developed by Brownsey and Maples (1966) is often used. The technique, however, is not free from complications (Simmons and Thompson 1968).

### 3.3. *Stress, sheet flatness and transformer noise*

The sensitivity of the magnetostriction of magnetic materials to stress is well known. In silicon-iron, the stress that has the most marked effect on the longitudinal magnetostriction is a compressive stress in the plane of the lamination and in the direction of the flux. The variation in magnetostriction with both compressive and tensile stress is shown in figure 4.

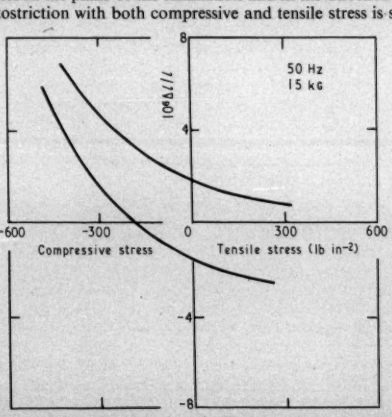



Figure 4. The magnetostriction of grain-oriented silicon-iron increases with compressive stress and decreases with tensile stress. In general, experimental curves fall within the boundary curves and have a similar shape.

The magnetostriction-stress curve is not unique, considerable differences are found from one sample to the next, although the manner of the variation is similar: a tensile stress causes a reduction in the magnetostriction, a compressive stress causes a marked increase in the magnetostriction.

The effect of stress on the magnitude of the magnetostriction is of considerable importance in the construction of large transformers (Wilkins and Thompson 1962). The laminations are never perfectly flat, so that when these laminations are built into a core detrimental stresses along the direction of the flux are produced, even when the sheets are considered flat by general metallurgical standards. Indeed, a basic and troublesome problem is to define and measure sheet flatness. Physical measurement of wave size would be tedious and difficult and, ideally, it is the average stress produced on flattening the laminations that is required. One method that has gained general acceptance is shown in figure 5. A stack is placed on a substantial bed plate and over the laminations is placed an aluminium plate attached to a flexible rubber gasket. The aluminium plates and rubber edge seal form an

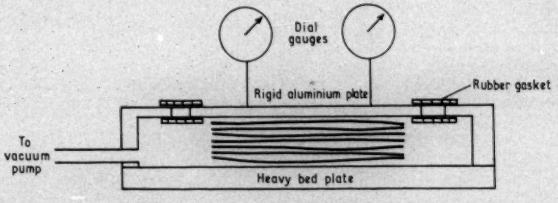



Figure 5. One simple means of comparing the sheet flatness of batches of laminations.

airtight chamber and the gasket allows the upper plate to move freely. As air is exhausted from the chamber, the differential air pressure on the aluminium plates gradually flattens the laminations, and the flattening is measured by dial gauges. A curve of load against deflection is plotted and the 'area under the curve' represents the work done in compressing the plates, which is equal to the mechanical strain energy stored in the laminations, from which the stress can be determined. In order to eliminate frictional effects the average of the curves, with both increasing and decreasing load, is taken. A pressure of 5 lb $in^{-2}$ is sufficient to produce almost complete flattening.

### 3.4. *Harmonics and transformer noise*

In general, the magnetostriction is taken to be the peak magnetostriction under ac conditions. Whilst this is reasonable, in many cases both the way in which the magnetostriction varies with the flux density (the well-known 'butterfly loop') and the magnitude of the harmonics of the 100 Hz fundamental are important. These may be small in magnitude, but very significant in the production of noise (Thompson and Wilkins 1963). In some power transformers it is the second harmonic that contributes primarily to the total noise; in others it may be the fourth or fifth harmonic. The influence of stress on the harmonic content of the magnetostriction loop, and its relevance to transformer noise, has been discussed by Holt (1967). His investigation showed that a wide variation in stress sensitivity of the magnitude of the harmonics exists between various grades of materials, and that these depend markedly on the type of coating and the nature of the heat treatment used in producing the sheets.

## 4. Possible new materials

### 4.1. *Introduction*

Over the years, with advances in metallurgical practice, the quality of grain-oriented silicon-iron has improved markedly. This process will undoubtedly continue; indeed, Gettings (1967) has recently discussed the question, "Is 3½% Si-oriented core steel the ultimate?" However, a number of developments have been taking place which could bring, potentially, even greater improvements. These are (*a*) the successful production on a laboratory scale of a different form of grain-orientation in 3½% silicon-iron, (*b*) the evolution of a technique for producing 6½% silicon-iron in sheet form with a suitable grain-orientation and (*c*) the development of entirely new alloys.

### 4.2. *A new oriented form of silicon-iron*

In 1957, Assmus *et al.* described how another orientation could be developed, with a [001] direction parallel to the rolling direction and the (100) plane parallel to the rolling plane, often referred to as the 'four-square' texture. The basic advantage of this material to the engineer is that it has almost equal magnetic properties in two directions at right angles in the plane of the sheet. In consequence, equipment has to be designed to take advantage of this factor, but in most cases the economic saving would correspond to only about £20 per ton of silicon-iron used (Thompson 1963). Unfortunately, the metallurgical process by which the new form of material was initially produced, when transferred to bulk production, is unlikely to be competitive with the present Goss process. However, as a result of further recent development work, it may be possible to modify the process and relax some of the strict processing conditions.

### 4.3. 6½ *wt% silicon-iron*

Adding silicon to iron increases the resistivity of the alloy, thus reducing the power loss of the laminations. The practical limit to the amount of silicon that can be added is about 3½ wt% silicon, above which the material becomes brittle and the necessary cold-reduction stage of the normal process is no longer possible. It would be desirable to produce 6½% silicon-iron, since it has been reported on a number of occasions that zero magnetostriction and minimum loss occur at this composition. Unfortunately, the increase in silicon content brings with it a reduction in the magnetic saturation induction. An assessment of the possible advantages of 6½% silicon-iron in large power transformers has been reported

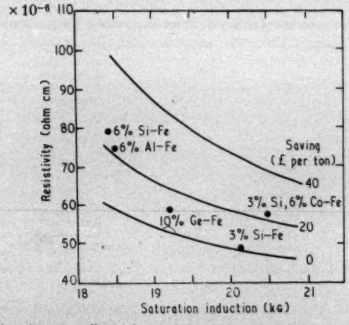



Figure 6. Curves showing the effect of resistivity (taken to represent power loss) and saturation induction on typical transformer costs—expressed as saving per ton of core material.

(Brown *et al.* 1964). The successful production of 6½% silicon-iron sheet on a small scale is evidently taking place in the U.S.A., since the performance of a 3-phase, 1250 V A static inverter has contrasted very favourably with the performance of equivalent inverters made from other materials, such as cube-textured 3½% silicon-iron laminations (Kernick *et al.* 1967).

4.4. *Other possible alloys*

Iron-aluminium has attracted attention over the years and it has been reported to have good magnetic properties (Nachman and Buehler 1954). However, using the Goss process, it would appear that special precautions are needed at the rolling and annealing stages. Aluminium oxidizes quite readily and, during rolling, small particles of the oxide usually are forced into the bulk material, with subsequent deterioration in properties. The problems associated with the production of oriented iron-aluminium are being restudied by Cahn, (Gardner and Cahn 1966), particularly the possibility of developing the orientation by the surface-energy process.

Any improvement in properties of new alloys can be expressed in economic terms for any specific usage by the engineer. In the particular case of a power transformer, the initial assessment of this potential of new binary alloys can be determined by evaluating the effect of the resistivity and saturation induction on transformer costs—expressed as a saving in £ per ton of core material (figure 6). There have been suggestions that, in ternary alloys, the additions might have greater beneficial effects, owing to interactions, that either of the two additions on their own. This is quite possible, since such interactions are known in other spheres. The development of oriented forms of any new alloys, which are competitive in price, lies in the province of the metallurgist.

## II. High permeability materials

### 5. Introduction

Some engineering applications require materials of very high permeability and low power loss, particularly for use at high frequency. The attainment of the desirable properties with a high degree of consistency is of prime importance, whereas the cost of the finished material is relatively unimportant. Thus sophisticated and expensive techniques may be used during their manufacture.

Most materials are based on the binary nickel-iron alloys containing 35–100 at.% nickel, with the addition of various elements such as molybdenum and copper. The name 'permalloy' is commonly used to describe the alloys.

The practical means of achieving high permeability were discovered by earlier workers (1923) in the field, and although today the treatments are basically similar, the exact mechanism whereby these heat treatments etc. control the properties is still subject to much discussion. Elmen (1929) recognized the importance of producing material of the highest, controlled, purity, and originated the 'double heat' treatment, now known as the permalloy treatment, that gives maximum permeability in alloys around 78% nickel.

Another important feature was discovered by Dillinger and Bozorth (1935), who heat treated the alloys in a magnetic field; a process referred to as magnetic annealing. By applying a field of about 1 Oe from a temperature of 500 °C, the maximum permeability of a 78 permalloy could be increased to above 100 000, which is the value obtainable by a 'double heat treatment'.

Over the last few years, an understanding of the factors that control the magnetic characteristics has developed. It was originally thought that high permeabilities could be obtained only in alloys within a narrow range of compositions, but this is no longer considered to be correct; indeed, good properties can be obtained in ternary and quaternary alloys containing copper and molybdenum with treatments that give poor properties in the normal binary alloys. Basically, high permeabilities are considered to arise from the combination of low magnetostriction (the interplay of magnetostriction and stress) and low anisotropy.

Before discussing recent experimental work, it is necessary to mention briefly a number of topics.

## 6. Some theoretical considerations

### 6.1. *Anisotropy and magnetostriction*

In alloys in which the anisotropy is very small, rotational processes could take precedence over wall movements. Chikazumi (1964) points out that, using a model based on coherent rotation magnetization, for a permeability of 10 000 the anisotropy needs to be as low as about 300 erg cm$^{-3}$, from which he concludes that coherent rotation is hardly likely to occur. It is more probable that displacement of broad walls, or incoherent rotation, is involved when the magnetization changes.

Both magnetostriction and strain, and their interplay, have some part in determining the permeability of nickel-iron alloys. It has been suggested that the combined strain energy arising from internal stress and magnetostriction, together with magnetocrystalline energy, give a minimum value of the total energy at a composition of 78·5% Ni in binary alloys. Lewis (1964) has considered the conditions necessary to make the total energy $E$ a minimum. He introduces the expression $E''(\theta_0)$ to represent the rotational stability of the magnetization vector. If the internal stresses were zero, the stability would be lowest and the permeability highest at 75% Ni, where the anisotropy constant is $K_1=0$. Since, experimentally, the permeability is highest at 78·5% Ni, he suggested that the stability is lowest there and that the internal stress has a maximum value $\sigma_m$, given by

$$|(\lambda_{100}-\lambda_{111})\sigma_m| \sim |K_1|.$$

Using values of $K_1$, $\lambda_{100}$ and $\lambda_{111}$ for the material,

$$E_c''(\theta_0) \sim K_1 = 1700 \text{ erg cm}^{-3}$$

and

$$\sigma_m \sim 1{\cdot}7 \times 10^{-8} \text{ dyn cm}^{-2}.$$

An expression for the initial permeability of a polycrystalline specimen in which only rotation occurs has been derived by Lewis and Street (1958). This is

$$\mu_0 = \frac{8\pi I_s^2}{3E_c''(\theta_0)}$$

which yields $\mu_0=3500$, on substituting suitable values for $I_s$ and $E_c''(\theta_0)$. This value is considerably less than the value normally associated with the binary 78·5 Ni–Fe alloy, i.e. 10 000. Thus it would appear that domain-boundary movements must contribute to the permeability.

### 6.2. *Magnetic annealing and atomic ordering*

The basic effect of cooling in a magnetic field is to introduce a uniaxial anisotropy of a few thousand erg cm$^{-3}$ in the direction of the field, which has two simple magnetic consequences. Firstly, the domain structure becomes one of parallel and anti-parallel domains separated by 180° walls lying parallel to the direction of the annealing field and, secondly, a very rectangular hysteresis loop, when measured in the direction of the annealing field, is produced. The most satisfactory theory to date is that of 'directional ordering', developed by Néel (1954) and Taniguchi and Yamamoto (1954, 1956).

Both models require a preferential distribution of unlike atoms pairs and yield similar predictions. As a consequence of the ordered structure, a uniaxial anisotropy is superimposed on the normal crystal anisotropy, which has a complicated angular dependence except in certain simple configurations. In general, the position of the energy minimum of the 'induced anisotropy' will not coincide with the direction of the annealing field. This feature is shown by the experimental results of Chikazumi (1956) in figure 7, where the angle

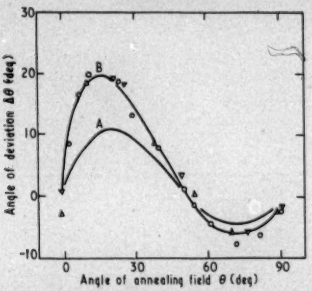



Figure 7. The variation of $\Delta\theta$ (the angle between the direction of the annealing field and the easy axis of the induced anisotropy) with the crystallographic orientation of the annealing field. Rates of cooling are: ○ 14 degC per min; △ 5 degC per min; ▽ 1 degC per min.

of deviation, $\Delta\theta$, of the minimum point from the direction of the annealing field, is plotted as a function of the direction of the annealing field. Although, qualitatively, the theoretical curve A agrees fairly well with the experimental curve B, there is a quantitative difference which has to be resolved.

Magnetic annealing should occur in all alloys where the activation energy at the Curie temperature is greater than that required for atomic migration.

### 7. Recent work on alloys around 80 Ni–20 Fe

A considerable improvement in these binary alloys can be achieved by an addition of molybdenum and/or copper of about 5%. A British Mumetal might have the composition 5% Cu, 4% Mo, 77% Ni and 14% Fe, but an American 4–79 permalloy of a similar type would contain no copper. The addition of molybdenum to the binary alloy can increase

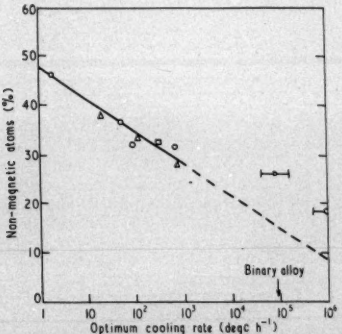



Figure 8. The change of optimum cooling rate with percentage of non-magnetic atoms (those carrying no resultant moment) in some Ni–Fe alloys (Enoch and Fudge 1966). ○ ternary; △ quaternary; □ quinary.

the initial permeability by a factor of ten; it increases the resistivity and simplifies the heat-treatment considerably.

The earlier work of Richards *et al.* (1957) has been extended by Enoch and Fudge (1966), whose alloys were prepared in the same way by powder metallurgy. They concluded from their work that a complex nickel–iron alloy containing molybdenum and/or copper will have a high initial permeability if two conditions are satisfied: (*a*) the ratio of magnetic nickel atoms is the same in the complex alloy as in the optimum binary alloy and (*b*) the degree of short-range order involving iron atoms and magnetic nickel atoms is the same as that in the optimum binary alloy. The increase in initial permeability over that of the best binary alloys was considered to be due to a further reduction in the already small values of magneto-striction and anisotropy. The percentage of non-ferromagnetic atoms present in an alloy, and the optimum cooling rate, were related experimentally, as in figure 8. The increase in the duration of the final heat treatment for complex alloys was thought to be due to the presence of non-ferromagnetic atoms hindering the short-range ordering process. This process becomes more sluggish as a result of the lowering of the long-range order–disorder temperature. In certain alloys some long-range order appears to be necessary to provide the optimum amount of short-range order. The upper limit of the value of the initial permeability was considered to be set by the stress arising from interactions between ordered and disordered regions, which have differing coefficients of expansion.

Similar results on the relationship between the optimum cooling rate and the amounts of various added elements have been obtained by Lykens (1966) for a vacuum-melted 4–79 molybdenum permalloy. Lykens also measured the initial permeability as a function of the amount of alloying addition in the range 0–1 wt% for Cr, Nb, Ti, W, Ta and Al. The initial permeability passed through a peak at about 0·6% for the three elements Cr, Nb and Ti, but the other three had no appreciable effect on either the optimum cooling rate or the initial permeability.

A series of vacuum-melted commercial alloys of the Mumetal type, containing 11·5–15·6% Fe, 4% Mo, 5% Cu (balance Ni), have been studied extensively by Scholefield *et al.* (1967). The effects of composition and variations in heat treatment on the magnetostriction and initial permeability were investigated. Initial permeabilities greater than 100 000 could be obtained throughout most of the range showing, as with the sintered alloys of Enoch and Fudge, that the permeability was not critically dependent on the composition. The bake temperature for peak initial permeability was almost linearly dependent on the Fe content of the binary alloys. Figure 9 shows how the initial permeability depended on both the bake temperature and quench rate for a quaternary alloy. In view of the possible influence of the magnitude of the saturation magnetostriction on initial permeability, this parameter was measured in some detail. However, it was concluded that the saturation magnetostriction was of no consequence as regards the change of initial permeability with heat treatment. They considered that the high initial permeabilities were achieved when the heat treatments were such as to produce the necessary critical degree of short-range order to give a low anisotropy constant. With several of the alloys, the optimum bake treatment was at a temperature above that for the long-range order–disorder transition. Above the order–disorder temperature (about 500 °C) long-range order falls rapidly to zero, but short-range order exists up to about 600 °C, the formation taking place in an extremely short period.

There seems to be general agreement that it is the degree of short-range order that is the governing factor controlling the permeability. For example, Enoch and Winterborn (1967) have measured the first anisotropy constant at room temperature of a 77% Ni, 5% Cu, 4% Mo, and 14% Fe alloy, and shown that the highest initial permeability and lowest coercive force occur with a vanishingly small anisotropy associated with a critical amount of order. The limit to the value of the permeability appears to be set by the magneto-elastic anisotropy arising from internal stresses. An alloy of the same composition has been studied by Jackson and Lee (1966), who point out that the magnetostriction constant is the one parameter most likely to be critically dependent upon short-range magnetic interactions. Consequently, they measure this parameter and the initial permeability $\mu_0$ at room

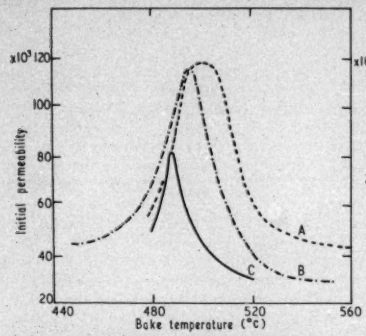



Figure 9. Initial permeability as a function of bake temperature and quench rate for a quaternary alloy. A quench rate 50 degC per min; B quench rate 150 degC per min; C quench rate 2500 degC per min.

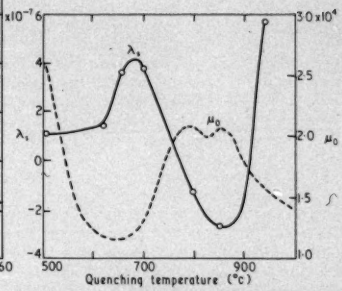



Figure 10. The magnetostriction constant $\lambda_s$ of a 77% Ni alloy measured at room temperature as a function of the quenching temperature, $t_Q$ (Jackson and Lee 1966).

temperature as a function of the quenching temperature $t_Q$, as shown in figure 10. Their general results were ascribed to short-range order and no evidence could be found of long-range ordering.

In general, most of the work is carried out on material in the thickness range $2\text{–}4\times10^{-3}$ in. Recently, however, Enoch and Winterborn have shown that, for the material below $3\times10^{-3}$ in thick, a thickness-dependent magnetostatic contribution to the wall energy becomes important.

The presence of oxygen in the material, arising from early stages in the production of the alloys, can have a marked effect on its final properties. Nesbitt *et al.* (1965) have established a connection between impurity faults on the (111) planes caused by oxygen, and the ability to respond to heat treatment in a magnetic field. The work suggested that oxygen inhibits long-range order and may aid the formation of like atom pairs.

## 8. Recent work on 50 Ni–50 Fe

An interesting new magnetic alloy of the 50 Ni–50 Fe type called Satmumetal (Dennis 1966) has a maximum permeability equal to that of 80 Ni–20 Fe types and a much higher saturation flux density, because of the higher basic iron content, which is desirable in a number of engineering applications. Typical properties are reported to be: $\mu_0=35\,000$, $\mu_{max}=240\,000$, $B_{sat}=15\,000$ G and $H_c=0{\cdot}025$ Oe. This material is available in 0·004 in strip.

These improved properties can be produced by annealing below the Curie temperature in a magnetic field (Kang *et al.* 1967). The degree of response to magnetic annealing was observed to be dependent on the nature of the secondary recrystallized textures, of which two kinds were observed. However, only one of these, where the central cube pole of a cube texture was rotated around the [100] axis parallel to the rolling direction, showed good response to magnetic annealing. Magnetostriction measurements were interpreted as showing that a uniaxial anisotropy was developed by short-range ordering of like atom pairs, which in turn produced a more favourable alignment of 180° domain walls.

## III. Thin ferromagnetic films

### 9. Introduction

The rapid development over the last ten years of digital computers has led to the detailed study of thin ferromagnetic films. The two main functions that occur within a computer are the storage and processing of information or data. The storage of information can be achieved in a variety of ways; essentially, any system that has two stable states (bistable) has the potential to be of use.

In all computers it is desirable to have the maximum speed of operation and the maximum storage capacity, but for economic reasons some compromise has to be made between these conflicting requirements, depending on the particular type of market at which the computer is aimed. Since not all the data needs to be immediately accessible, some information can be stored cheaply in a serial manner on magnetic tape or drums. When access at any time ('random access') is required, thin magnetic films can be used (figure 11). Mixed stores reduce the cost per unit capacity.

Although much work is being carried out on serial stores, the major effort is concentrated on schemes utilizing the high-speed switching characteristics of thin ferromagnetic films. Such stores should fulfil the basic requirements for a memory store, which are: (*a*) that the device must have two stable states, (*b*) that it is consistent in properties and cheap to make, preferably by batch fabrication, and (*c*) that it is capable of being switched in a few nanoseconds, to match the operating times of logic elements. The fact that a piece of magnetic material can be magnetized in either of two opposite directions fulfils the first condition and, potentially, processes such as evaporation or electrodeposition satisfy the second. Fast switching times in thin films arise because there is a critical thickness, below which the film

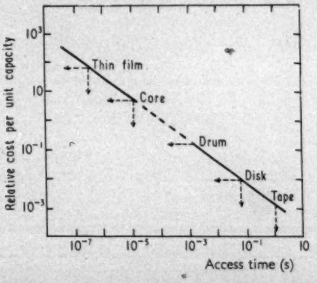



Figure 11. The relationship between the relative cost per unit capacity and the access time for various magnetic stores. Thin dotted lines indicate probable trends.

acts like a single domain particle, wherein reversals of magnetization proceed by coherent rotation of the spins and not by the slower process of domain-wall motion.

When the behaviour of magnetic thin films was studied in detail, it soon became apparent that films had novel physical properties, and much effort in recent years has been devoted to understanding these properties and developing the engineering aspects of thin film memories.

### 10. Some basic properties of magnetic films

That two main types of domain walls exist is well known; in thicker materials conventional Bloch walls are present, but below a certain thickness Néel (1955) walls are found. The transition from a Bloch wall to a Néel wall does not take place suddenly, and systems involving 'cross-tie' walls (Huber *et al.* 1958) have been found.

However, the concept and theory of single-domain particles (or discrete volumes), as developed by Stoner and Wohlfarth (1948), has had important implications in magnetic materials. In particles of less than a certain size, the existence of domain-walls is energetically unfavourable. Such particles are called single-domain particles. Magnetization changes take place by the coherent rotation of the spins. If the particle has some form of anisotropy then, as the field is changed cyclically in magnitude, the magnetization can trace out a $B$–$H$ loop whose 'squareness' depends on the angle between the 'easy' direction of magnetization and the direction of the applied field. In the case of particles, the major source of anisotropy can arise from an elongated shape.

The relevance of the Stoner–Wohlfarth theory to thin films is, firstly, that coherent rotation of spins is a much faster mode of changing the magnetization than by domain-wall motion and, secondly, a very square $B$–$H$ loop should be obtained in an anisotropic film. The anisotropy field $H_K$ is the field required to rotate the magnetization coherently from the easy direction to the hard direction against a uniaxial anisotropy. In the simplest case

$$H_K = \frac{2K}{I}$$

where $K$ is the uniaxial anisotropy and $I$ the intensity of magnetization.

In a permalloy film, this anisotropy may be as little as 1000 erg cm$^{-3}$, so that an applied field of 2·3 Oe is capable of rotating the magnetization through 90°.

Unfortunately, with practical thin ferromagnetic films, a number of disturbing factors such as dispersion, skew and creep influence the performance of the film. These will be discussed later.

## 11. Recent developments in the preparation of thin films

A number of techniques have been used in the preparation of thin ferromagnetic films which have had, as twin major aims, reproducibility of characteristics and low cost.

### 11.1. *Vacuum evaporation*

It soon became apparent that the properties of films prepared by vacuum evaporation depended markedly on a number of factors, such as the rate of evaporation, the nature and temperature of the substrate, the degree of vacuum present, etc. Typical of the early work is that reported by Blois (1955). Giedd and Perkins (1960) measured and controlled the rate of evaporation by means of a modified ionization gauge placed in the vapour stream. The measurement of the rate of deposition and film thickness has presented problems. One technique (Oberg and Lingenso 1959) involves noting the change of frequency of a crystal of quartz in a resonant circuit, as deposition on to the crystal proceeds.

Of course, during any vacuum evaporation process the degree of vacuum and the nature of the residual gases are generally important. Paul and Hanson (1966) concluded that $N_2$, $H_2$, $O_2$ and $H_2O$ had no noticeable effect on the anisotropy for pressures below $10^{-5}$ torr. However, Young (1967) has found that films of Ni–Fe (82 : 18), prepared at $10^{-9}$ torr, are more reproducible and consistent with improved characteristics than those formed at $10^{-6}$ torr in the same vacuum system. The distributions of the magnitude dispersion in $H_K$ (denoted by $\Delta H_K$), and of the angular dispersion $\alpha$, are shown for a large number of films, in figures 12(*a*) and 12(*b*). In contrast to Paul and Hanson (1966), Young's limited work on the introduction of known gases has shown that $O_2$ has a marked effect on the magnetic properties.

The direction and magnitude of the anisotropy may change from place to place in a film, giving rise to magnetic dispersion, the symbol $\pm\alpha_m$ denotes the maximum angular value of the dispersion from the 'easy' axis. In practice, because of the techniques of measurement, angular dispersion is not obtained directly, but rather the dispersion of the direction of magnetization as the element is rotated through 90° from the 'easy' axis by the application of a transverse magnetic field. This gives rise to an effective angular dispersion which, for computer use, should not exceed 4° (Petschauer 1965).

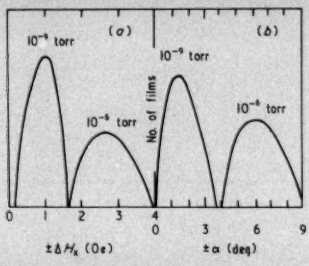



Figure 12. The distribution of (*a*) the magnitude dispersion in $H_K(\Delta H_K)$ and (*b*) angular dispersion ($\alpha$) for a large number of films prepared under two different vacuum conditions (Young 1967).

11.2. *Electron beam bombardment and flash evaporation*

Vacuum evaporation is essentially a simple technique in which the crucible (or a source) is heated. However, changes in the composition of films arise from fractionation, even if the source is large in comparison with the amount of material being evaporated. Furthermore, a number of workers (Astwood and Prutton 1963) have indicated that impurities from the crucible contaminate the melt and can have a marked effect on the final properties of the film. In contrast with this, Wilts (1965) has reported that, for a large range of compositions and crucible materials, no detectable change that was attributable to crucible material could be observed. Nevertheless, attempts have been made to avoid compositional changes during evaporation and to limit, as far as possible, the entry of contaminants into the melt.

In electron beam bombardment (Heavens *et al.* 1959), a beam of electrons is made to impinge directly on to the surface from which evaporation is required. This increases the efficiency of the evaporation, but, still more important, contamination of both the melt and the evaporated material by the crucible is reduced. The risk of contamination can be further reduced by eliminating the crucible entirely and placing the ingot on a water-cooled hearth.

A flash evaporation technique has been used by Heritage (1967), in which each constituent of the alloy is evaporated from a hot strip. It was pointed out that, with films of Ni–Fe (81 : 19) prepared by vacuum evaporation in an aligning magnetic field, the magnitude of the anisotropy field ($H_K$) was greater than is desirable for a high-speed store. The anisotropy induced by the field has been attributed to Fe–Fe atomic pairings (Taniguchi 1955) and Heritage considered that the addition of a third element would be beneficial. He concluded that films of Ni–Fe–Ge offer some advantages over binary Ni–Fe films and that flash evaporation gives improved compositional control.

11.3. *Cathodic sputtering*

In ionic or cathodic sputtering, a target is eroded by bombardment with positive ions which are usually obtained from a gaseous plasma. Sputtering should offer some advantages over vacuum evaporation, for elements having very different vapour pressures, and lead to increased uniformity in batch production (Spoliotis 1967). Recently, a thorough investigation of sputtering as a technique has been carried out by Mayer (1966). He concluded that, in general, sputtered magnetic films resemble evaporated magnetic films in that the gross parameters vary in the same manner with processing variables. Figure 13 shows the reproducibility of $H_K$ and thickness which Mayer obtained for 65 samples. It was concluded that such films exhibit excellent pulse-switching characteristics and result in high quality thin-film elements.

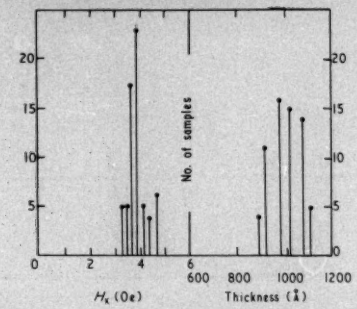



Figure 13. The distribution of $H_K$ and thickness for 65 samples deposited at 300°C (Mayer 1966).

11.4. *Electrodeposition*

Electrodeposition has attracted some attention recently (Girard 1962, Wolf 1967), since it provides an easy means of introducing impurity atoms into thin films. However, most of the work seems to be concentrated on studying various basic physical phenomena that occur in such films. The effect of sulphur on the characteristics of Ni–Fe electrodeposited films has been studied by Luborsky (1968), who discussed the behaviour of vacuum deposited and electrodeposited films. Both have similar annealing characteristics.

11.5. *Preparation of hard magnetic films*

It is, perhaps, relevant at this stage to mention present research into the preparation of hard magnetic films for high-density recording. Both electrodeposition and chemical deposition have been used to prepare such films. For a high recording density, a high ratio of coercive force to thickness for a given remanent magnetization is required. It is thought generally that the high coercive force observed with cobalt–phosphorous films is a result of the switching behaviour of fine particles, arising from the isolation of cobalt crystallites or particles, by phosphorous. Details of the properties of Co–P films have been reported by Blackie *et al.* (1967), and for Co–Zn–P films by Fisher (1966); the addition of zinc appears to increase the degree of isolation of the cobalt particles.

## 12. Recent device and engineering developments

12.1. *Early devices*

Some of the most successful computer stores have used small ferrite (usually manganese magnesium ferrite) ring cores as the basic element. Such cores may have an outside diameter of 0·050 mm and an inside diameter of 0·030 mm and, when suitably driven, may have a switching time $\simeq 10^{-6}$ s. Their organization into an array and the method used for writing-in and reading-out information is well known. Read-out, or interrogation, is destructive in the sense that the information needs to be rewritten to be retained in the memory; such devices have 'destructive read out' (D.R.O.). Devices which can be interrogated without the loss of their information, 'non-destructive read-out' (N.D.R.O.), offer obvious savings in access time, and devices such as the two-apertured transfluxor (Rajchman and Lo 1955, 1956) fall into this category.

Magnetic material in the form of wire has been used in devices named 'twistors' for random-access memories (Bradley 1960). These are constructed by wrapping a magnetic wire around an insulated non-magnetic copper wire in the form of a helix making an angle

of 45° with respect to the axis. A solenoid is wound over both. Combined currents through this solenoid and the wire produce a magnetizing force parallel to the helical wire. The magnetic wire is under longitudinal tension and, as a consequence, has a rectangular hysteresis loop. Continuous sheet thin-film memories have also been developed (Bean and Thiel 1966).

12.2. *Multi-layer film devices and some general design problems*

Even with the simplest magnetic planar arrays, departures from ideal rectangularity of the hysteresis loop give rise to induced voltages in the read-out windings from cores not being interrogated. These are called 'disturb' voltages. By linking the read wires in certain ways, some cancelling of such voltages is possible. However, other basic magnetic non-uniformities produce disturb-voltages, and the stability of an individual element depends on the design of the magnetic array; very strongly so in the case of a miniaturized multilayer device. Disturb-fields limit the achievable bit-density of a film.

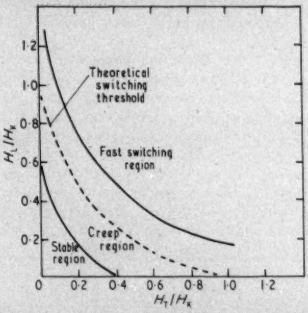



Figure 14. Graph showing the three regions of interest; defined by the various combinations of transverse ($H_T$) and longitudinal ($H_L$) fields which are applied to the film.

Domain-wall creeping is a process involving the slow motion or drift of domain walls. It occurs for particular combinations of applied fields, and reduces the disturb-threshold of a film. Repeated transitions parallel to a film's 'hard' axis in the presence of an 'easy' axis field, often give rise to slow motion or creeping of the domain boundaries. In figure 14, three regions should be noted (Petschauer 1965), defined by the various combinations of transverse and longitudinal fields applied to the film. In the stable region, any combination of transverse and longitudinal fields can be applied repeatedly without causing measurable loss of magnetization. One method of measuring creep has been described by Bean and Thiel. Creep effects are most pronounced in the thickness range 200–900 Å (Gagnon and Crowther 1965), which appears to be associated with cross-tie walls, whose number per unit area changes drastically in this range (figure 15, Methfessel *et al.* 1960, Hulyer 1964). In films with thickness below about 200 Å, creep has not been observed in unidirectional hard-axis films.

Unfortunately, in films of less than about 300 Å thickness the signal obtained on sensing is too low for most applications, and so Middlehoek and Wild (1967) have suggested the use of a composite film in which two films of magnetic material are separated by a thin non-magnetic layer such as silicon monoxide. Multi-layer films consisting of alternate layers of magnetic and non-magnetic layers are a simple extension of this idea. Siegle and Flur (1967) have shown that significantly improved results, compared with a conventional 600 Å

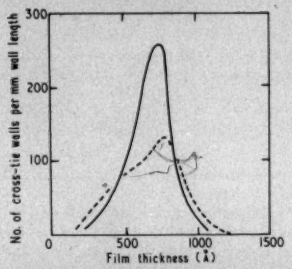



Figure 15. Creep effects appear to be associated with cross-tie walls whose number changes markedly with thickness in range 200–900 Å (full and broken curves by two different workers Methfessel *et al.* 1960 and Hulyer 1964).

film, can be obtained by using two 300 Å non-magnetostrictive films separated by a 1 $\mu$m layer of silicon monoxide. These thicknesses appear to give optimum results.

A pair of magnetostatically coupled films with their 'easy' axes at 90° to each other, exhibit biaxial anisotropy and, when an external bias magnetic field is applied along the easy direction of one of the films, only two states of the coupled system have minimum energy. A device consisting of electroplated magnetic layers, conductive layers and photo etched strip lines, which may be capable of development for specific purposes, has been fabricated by Wang (1968). A multi-layer film device, consisting of two nickel–iron films coupled through a non-magnetic (e.g. gold) film, has been investigated by Stein and Feldtkeller (1966) as a word-organized D.R.O. memory. The tolerance in such parameters as drive current was eased considerably, both by the effects of coupling and also by the use of a ferrite keeper.

### 12.3. *Some special devices*

Various ideas have lead to the development of novel devices, which have sought to overcome the drawbacks associated with the more conventional memories. It has been demonstrated that planar films of Ni–Fe, shaped like a diamond with the easy axis along the long diagonal, can be switched rotationally by the combination of two fields, $H_L = 2/3H_c$ and $H_T = 2/3H_K$, each of which is incapable of switching the film separately. It has been predicted by Barker *et al.* (1968) that an N.D.R.O. memory of $10^6$ bits with a 60 ns cycle time is possible.

A new mode of controlling magnetic domain propagation has been reported by Michaelis (1968), though it is not clear how this can best be used in a computer store.

An unusual sonic film memory has been suggested by Weinstein *et al.* (1968) which utilizes the strain sensitivity of a magnetostrictive thin film for the storage and retrieval of information. In this form of application it is essential to have maximum strain sensitivity, rather than the minimal strain sensitivity normally desired. Writing is obtained by the coincidence of a strain pulse and a magnetic field pulse. Non-destructive read-out is accomplished by subjecting the film to a scanning strain pulse. Onyshkevych (1968) has prepared films of Ni–Fe–Co (60 : 25 : 15) by evaporation from the melt, which have suitable characteristics.

The use of magneto-optics in computers has been developed in recent years and will be discussed briefly in the final section.

### 12.4. *Magneto-optics and computer memories*

With the usual sensing techniques, there is a lower limit to the amount of material that

must be switched in order to obtain an output of sufficient magnitude to overcome disturbances such as thermal noise in the sensing channel and varying demagnetizing fields. It follows that there is an upper limit to the useful bit density of any memory array of a given size. A method of sensing is therefore desirable in which the output does not depend on unit size or bit density alone. One technique which is being developed on these lines is magneto-optical sensing, in which the output depends on the intensity of the sensing light beam. If writing could be accomplished optically further advantages would accrue, hence the study of various optical schemes.

One scheme has been discussed by Smith (1967). This depends for its feasibility on a high-power (cm wavelength) laser, which could interrogate simultaneously all the words in a memory. N.D.R.O. could be achieved with a composite layer, consisting of a Ni–Co alloy of high coercive force as the storage film, and a permalloy film of low coercive force as the interrogation film since, under these conditions, the threshold for interrogation writing would be much less than that for memory writing.

## References


ASSMUS, F., DETENT, K., and IBE, G., 1957, *Z. Metalk.*, **48**, 344.
ASTWOOD, P., and PRUTTON, M., 1963, *Br. J. appl. Phys.*, **14**, 48.
BARKER, P. D., OREDSON, H. W., and TOROK, E. J., 1968, *J. appl. Phys.*, **39**, 1213.
BARRETT, W. F., BROWN, W., and HADFIELD, R. A., 1900, *Sci. Trans. R. Dubl. Suc.*, **1**, 67.
BEAN, W. R., and THIEL, F. L., 1966, *I.E.E.E. Trans. Magnetics*, **2**, 21.
BLACKIE, I., TRUMAN, N., and WALKER, P., 1967, *Proc. Conf. on Magnetic Materials* (London: I.E.E.), **33**, 207.
BLOIS, M. S., 1955, *J. appl. Phys.*, **26**, 975.
BOON, C. R., 1967, *Proc. Conf. on Magnetic Materials* (London: I.E.E.), **33**, 110.
BOON, C. R., and ROBEY, J. A., 1967, *Proc. Conf. on Magnetic Materials* (London: I.E.E.), **33**, 105.
BOON, C. R., and THOMPSON, J. E., 1965, *Proc. I.E.E.*, **112**, 2147.
BRADLEY, E. M., 1960, *Electronics*, **33**, 78.
BRAILSFORD, F., 1948, *J.I.E.E.*, **95**, 38.
BRAILSFORD, F., and ABU-EID, Z. H. M., 1963, *Proc. I.E.E.*, **110**, 751.
BRAILSFORD, F., and BURGESS, J. M., 1961, *Proc. I.E.E.*, Monograph 446 M.
BRAILSFORD, F., and FOGG, R., 1966, *Proc. I.E.E.*, **131**, 1562.
BRAILSFORD, F., and MAZZA, V. R., 1962, *Proc. I.E.E.*, **109A**, 173.
BROWN, D., HOLT, C., and THOMPSON, J. E., 1963, *Proc. I.E.E.*, **111**, 1933.
—— 1965, *Proc. I.E.E.*, **112**, 183.
BROWNSEY, K. M., and MAPLES, G. C., 1966, *Proc. I.E.E.*, **113**, 1859.
CHIKAZUMI, S., 1964, *Physics of Magnetism* (New York: J. Wiley), p. 264.
—— 1956, *J. Phys. Soc. Japan*, **11**, 551.
DENNIS, W. H., 1966, *Electl Rev., Lond.*, **178**, 372.
DILLINGER, J. F., and BOZORTH, R. M., 1935, *Physics*, **6**, 279.
ELMEN, G. W., 1959, *J. Franklin Inst.*, **207**, 583.
ENOCH, R. D., and FUDGE, A. D., 1966, *Br. J. appl. Phys.*, **17**, 623.
ENOCH, R. D., and WINTERBORN, Anne, 1967, *Br. J. appl. Phys.*, **18**, 1407.
—— 1967, *Proc. Conf. on Magnetic Materials* (London: I.E.E.), **33**, 123.
FISHER, R. D., 1966, *I.E.E.E. Trans. Magnetics*, **2**, 681.
GAGNON, G. P., and CROWTHER, T. S., 1965, *J. appl. Phys.*, **36**, 1112.
GARDNER, G. A., and CAHN, R. W., 1966, *J. Mater. Sci.*, **1**, 211.
GIEDD, G. R., and PERKINS, M. A., 1960, *Rev. sci. Instrum.*, **31**, 773.
GETTINGS, M. P., 1967, *J. appl. Phys.* (Supplement), **38**, 1074.
GIRARD, R., 1962, *J. appl. Phys.*, **38**, 1423.
GOSS, H. P., 1934, *U.S. Patent No.* 1,965,599.
HEAVENS, O. S., BROWN, M. M., and HINTON, V., 1959, *Vacuum*, **9**, 17.
HERITAGE, R. J., 1967, *Proc. Conf. on Magnetic Materials* (London: I.E.E.), **33**, 191.
HOLT, C., 1967, *Proc. Conf. on Magnetic Materials* (London: I.E.E.), **33**, 96.
HUBER, E. E., SMITH, D. O., and GOODENOUGH, J. B., 1958, *J. appl. Phys.*, **29**, 294.
HULYER, P. J., 1964, *Mullard Res. Labs., Rep.* 505.
JACKSON, R. C., and LEE, E. W., 1966, *J. Mater. Sci.*, **1**, 362.

KANG, I. K., SCHOLEFIELD, H. H., and MARTIN, A. P., 1967, *J. appl. Phys.*, **38**, 1178.
KERNICK, A., LANE, D. H., OGDEN, J. W., and PAVLOVIC, D. M., 1967, *J. appl. Phys.* (Supplement), **38**, 1087.
LEAK, D. A., and DUCKWORTH, W. E., 1963, *J. Iron Steel Inst.*, **201**, 588.
LEE, E. W., 1958, *I.E.E.*, Monograph **284M**.
LEWIS, B., 1964, *Br. J. appl. Phys.*, **15**, 407.
LEWIS, B., and STREET, R., 1958, *Proc. Phys. Soc.*, **72**, 604.
LUBORSKY, F. E., 1968, *I.E.E.E. Trans. Magnetics*, **2**, 681.
LYKENS, A. A., 1966, *J. appl. Phys.* (Supplement), **37**, 1204.
MAYER, W. N., 1966, *I.E.E.E. Trans. Magnetics*, **2**, 166.
METHFESSEL, S., MIDDELHOEK, S., and THOMAS, H., 1960, *IBM J. Res. Dev.*, **4**, 96.
MICHAELIS, P. C., 1968, *J. appl. Phys.*, **39**, 1224.
MIDDELHOEK, S., and WILD, D., 1967, *IBM J. Res. Dev.*
NACHMAN, J. F., and BUEHLER, J., 1954, *J. appl. Phys.*, **25**, 307.
NÉEL, L., 1954, *J. Phys. Radium, Paris*, **15**, 525.
—— 1955, *C.r. Hebd. Séanc. Acad. Sci., Paris*, **241**, 533.
NESBITT, E. A., BATTERMAN, B. W., FULLERTON, L. D., and WILLIAMS, A. J., 1965, *J. appl. Phys.* (Supplement), **36**, 1235.
OBERG, P., and LINGENSO, J., 1959, *Rev. sci. Instrum.*, **30**, 1053.
ONSHKEVYCH, L. S., 1968, *J. appl. Phys.*, **39**, 1211.
OVERSHOTT, K. J., PREECE, I., and THOMPSON, J. E., 1968, *Proc. I.E.E.*, **115**, 1840.
PAUL, M. C., and HANSON, M. M., 1966, *J. appl. Phys.*, **30**, 3743.
PETSCHAUER, R. J., 1965, *I.E.E.E. Trans. Magnetics*, **1**, 185.
PRY, R. H., and BEAN, C. P., 1958, *J. appl. Phys.*, **29**, 532.
RAJCHMAN, J. A., and LO, A. W., 1955, *RCA Rev.*, **16**, 303.
—— 1956, *Proc. Inst. Radio Engrs*, **44**, 321.
RICHARDS, C. E., WALKER, E. Y., and LYNCH, A. C., 1957, *Proc. I.E.E.*, **13**, 104.
SCHOLEFIELD, H. H., MAJOR, R. V., GIBSON, B., and MARTIN, A. P., 1967, *Br. J. appl. Phys.*, **18**, 41.
SIEGLE, W. T., and FLUR, B. L., 1967, *I.E.E.E. Trans. Magnetics*, **3**, 484.
SIMMONS, G. H., and THOMPSON, J. E., 1968, *Proc. I.E.E.*, **115**, 1835.
SMITH, D. O., 1976, *I.E.E.E., Trans. Magnetics*, **3**, 433.
SPOLIOTIS, D. E., 1967, *J. appl. Phys.*, **38**, 1207.
STEIN, K. U., and FELDTKELLER, E., 1966, *I.E.E.E. Trans. Magnetics*, **2**, 484.
STONER, E. C., and WOHLFARTH, E. P., 1948, *Phil. Trans. R. Soc.* A, **1240**, 74.
TANIGUCHI, S., and YAMAMOTO, M., 1954, *Sci. Rep. Res. Insts. Tōhoku Univ.*, **6**, 30.
—— 1956, *J. Phys. Soc. Japan*, **11**, 604.
TANIGUCHI, S., 1955, *Sci. Rep. Res. Insts. Tōhoku Univ.* A, **7**, 267.
THOMPSON, J. E., and WILKINS, J. T., 1963, *Electl Times*, **144**, 727.
THOMPSON, J. E., 1963, *Electl Times*, **144**, 877.
WANG, C. P., 1968, *J. appl. Phys.*, **39**, 1220.
WEINSTEIN, H., ONYSHKEVYCH, L. S., KARSTAD, K., and SHAHBENDER, R., 1967, *R.C.A. Rev.*, **28**, 317.
WILKINS, J. T., and THOMPSON, J. E., 1962, *Electl Times*, **141**, 233.
WILLIAMS, H. J., SHOCKLEY, W., and KITTEL, C., 1950, *Phys. Rev.*, **80**, 1090.
WILTS, C. H., 1965, *Br. J. appl. Phys.*, **16**, 1775.
WOLF, T. W., 1967, *J. appl. Phys.*, **33**, 1152.
YOUNG, A. S., 1967, *Proc. Conf. on Magnetic Materials* (London: I.E.E.), **33**, 197.

C. L. BOLTZ

# Back to the old Trend

## Arguments and background to the Green Paper

The Government's leaflet: *Industrial Research and Development in Government Laboratories*, otherwise known as the Green Paper, was published on 16th January 1970. Its importance to thousands of physicists and other scientists cannot be ignored, both for their own futures and the future of Britain's technological effort in relation to economic growth, a matter of concern to all research physicists. Its consideration is urgent, for though the Minister of Technology says in his foreword that the proposals merit wide public debate before decisions are taken by the Government, the fact is that a General Election is possible in October, and between now and then comes the long parliamentary recess. If use is to be made of the Government's present majority to ensure the acceptance of its decision, this must be made public by about June, leaving less than two months for the wide public debate. Indeed Lord Delacourt-Smith said in the House of Lords on 5th February last that the Government aimed to be in a position to give a clear indication of policy by the early summer. There is fear in some quarters that even less than two months will be available.

Before considering the proposals here it is appropriate to look back to October 1963, when a *Committee of Enquiry into the Organization of Civil Service* was published. It became known everywhere as the Trend Report, and scientists can be forgiven if they have forgotten what it was about, for there has been a series of publications known under the names of the chairmen—Jackson, Swann, Dainton, and many others. Its main recommendations relevant to the present situation were that the then Department of Scientific and Industrial Research (DSIR) should be dissolved and its functions and responsibilities distributed between three new research agencies. These were to be the Science Research Council, a new Natural Resources Research Council, and a new autonomous Industrial Research and Development Authority (IRDA).

This new Authority was to have several divisions and its function was to be to promote industrial research and development. It was to take over the Research Associations and DSIR research establishments and the National Research Development Corporation (NRDC) and it foresaw that if there came any change in the matter of defence policy and security in the Atomic Energy Authority then 'it would be natural to consider whether any of the Authority's civil functions should be combined with those of IRDA'. Within an hour the NRDC issued a statement and Sir William Black (now Lord Black), its chairman, said he had no relish for serving on a sub-committee of a sub-committee. Even Sir Nevill Mott, chief of the Cavendish Laboratory, opening the National Physical Laboratory's new Basic Physics building in May 1964, commented forcibly that he had not found the world 'sales' dotted about the Trend Report. He criticized the implied divorce between 'useless' science and 'useful or dirty' science linked with industry.

He considered that the Basic Physics Division could make the same sort of contribution to the industrial side of a national research effort that the Bell Telephone and U.S. General Electric Company could so outstandingly do for American industry.

These comments were all the more impressive for coming from an eminent scientist who was a 'pure' physicist.

These criticisms and others had perhaps a little effect. The DSIR *was* split as advised in the Trend Report and the basic science was allotted to the Science Research Council. The natural-resources research was given to a new Natural Environment Research Council. But nothing was done about the proposed new IRDA. The NRDC, NPL, NEL and most others of the research establishments and Research Associations were soon brought under

the administrative umbrella of the new Ministry of Technology, created by the Government and finalized in the Science and Technology Act of 1965. An Office of Scientific and Technical Information was formed separately, though it had been one of the proposed divisions of the IRDA. MinTech had taken over five of the six divisions proposed for the IRDA but now instead of an independent organization there was a Ministry, a colossus that may have begun to feel perplexed and worried in face of the many problems with which it has to deal.

Of the chief establishments that were to have been taken into the IRDA, NRDC went on independently and, with improved financial arrangements and greater freedom and under its energetic managing director, John Duckworth, eventually broke even with a small surplus of £48 000 on its 1968–69 activities. It became a force in the interface between research-and-invention and industry-and-commerce. The new director of the NPL, John Dunworth, began to rationalize the organization with a view to practicality, pruning off old-established departments and research that he considered was in effect over. It went on from strength to strength, earning about a third of a million pounds last year from industry (excluding earnings from the Government) despite the fact that much of its work is not intended for repayment. The National Engineering Laboratory also has achieved a world-wide reputation and has begun to build up as a regional centre of research for the Scottish engineering industry.

Thus, with more and more orientation towards practicality and exploitation, there seemed no immediate reason to resuscitate the proposal for a new Authority. The former Trend was, it seemed, forgotten. Yet it has appeared again, this time in the recent Green Paper. Why? The first and official reason is given in the Minister's foreword. 'Despite heavy national spending on research and development in Britain we have not profited fully from this investment for our rate of economic growth has been running at a lower level than that of many of our competitors.' (To this Lord Balogh retorted in the House of Lords that we have had very low returns from *any* investment in this country.) Mr Benn then goes on to mention the redirection of Governmental research effort to assist industry, the reductions of staff and reshapings of work programmes, and finally the alteration in size and shape of the U.K. Atomic Energy Authority (AEA), especially now that two of its activities are to be turned into separate companies, one for the production and sale and reprocessing of nuclear fuels and the other for the manufacture and sale of radioisotopes. 'The Government believes therefore that the time has come to consider a fundamental reorganization and a further redirection of the research and development resources of the Ministry of Technology and of the Atomic Energy Authority.'

The main reason for a fundamental reorganization now is thus seen to be the run-down of the AEA. This run-down really began in 1965 and was first reported in the press in October 1966, when Sir William Penney, Chairman of the AEA, said that he anticipated no major surgery though the scale of research and development would have to be reduced in the next few years. The cause of the run-down was that this enormous organization, which had reached a staff of nearly 41 000 in 1961, had served the purpose for which it was set up. This was twofold—to produce weapons and nuclear power, starting from scratch. The weaponry has been soft-pedalled and the nuclear-power side has reached a stage where nuclear power is established commercially and a smaller establishment is adequate and in any case the new nuclear-power consortia can cope with some development themselves. This fact of life, that if one races towards an objective and reaches it there is no need to run any further, constitutes a wry commentary on the prestigious way in which the Authority was built up, publicized, admired and envied. To-day, the further reduction of the establishment by the creation of competitive commercial organizations has left the remaining staff, conscious of the run-down, with low morale, and if anyone is insensitive enough not to take this seriously he knows nothing about scientific research and people. Many top-level scientists who sensed what was happening even before 1965 got away to posts in industry and universities and research elsewhere including overseas. It could almost be said that when Sir John Cockcroft left the AEA the morale was already declining. To-day some 3000 scientists (many of

them physicists) and engineers remain in the still-existing establishments (excluding AWRE concerned very much but not entirely with defence).

It must be appreciated that the AEA rouses strong feelings among outsiders. From the beginning it recruited scientists of the highest quality and they formed an élite. Atomic energy made news. Even Harwell, ostensibly for research only, with its wire fences and police and pass system maintained its secrecy and when a few of us were allowed to visit it first some twenty or so years ago we felt very privileged. (Capenhurst has even now never been visited by outsiders.) It is contended by some critics that all this created an arrogance among the élite. The AEA came to wield great power not always wisely used. This was evident, for example, in the affair of the nuclear ships in 1963–64 when the ill-fated Captain Atkins of Vickers criticized the AEA, which had acted as judge and plaintiff in considering rival systems up for consideration and then refused to face the critic on TV. Gossip among technologists claims that men of industry have found it difficult to get on with AEA scientists. These are all prejudices of course but they cannot be ignored. They give a foretaste of what may be said if the proposals of the Green Paper do get wide *public* debate. It is understandable if many people, even if eminent in technology, are less than dismayed about the slow death of the AEA. This discussion here may give some understanding of what is happening now and adds to the short history given in the Green Paper.

Let us go back to this and see what it actually says about the present and future. On the cover it says *A New Organization for the Seventies*; the image is surely here the message. Part I gives some history of the DSIR and other establishments in order to delineate a pattern of change. Part II describes the present scene and gives arguments for a change. Part III is entitled *A British Research and Development Corporation*, and it is largely this that any public debate will be about.

Part II underlines what has already been done in order to direct research towards industry and admits that scientific horizons beyond immediate programmes of work must be surveyed: so basic research will still be done in R and D organizations. (This wholeness of research was what Sir Nevill Mott was driving at in 1964.) The Green Paper also admits that there is now a new attitude towards industry and practicality among scientists, engineers and technicians in the various establishments. Indeed some of this Part II reads like a case against any further interference by reorganization. Then it goes on to say that many establishments are part of the Ministry's defence procurement organization; so no change in these is proposed. Here, it seems, is a real difficulty, for several of these establishments have made important advances in the civil field. The Royal Aircraft Establishment, for example, first produced the strong carbon fibres. The method was invented there. The Royal Radar Establishment has done notable work in electronics and semi-conductors. The Gunn effect was pioneered there. Many more examples could be given. All of this work is relevant to the whole vast question of civil research directed towards practical ends and industrial exploitation. In the dismissal of the several aviation research establishments as part of the new Corporation MinTech invites considerable criticism. The Green Paper goes on to state, without reasons, that the Research Associations will not be directly involved in any reorganization. It also argues that several of the Ministry's research stations are more governmental than industrial and omits several from the reorganization. They are the Laboratory of the Government Chemist (which does no work directly for industry), the Fire Research Station (though industry must surely be interested in firefighting), the Torry Research Station (which serves the fishing industry well) and the Water Pollution Research Laboratory (which does unique work for industry).

Five MinTech research establishments are thus left: NPL, NEL, Warren Spring Laboratory, Hydraulics Research Station, and Forest Products Research Laboratory. To these is to be added NRDC, the activities of which are, says the Green Paper, on a limited scale. Advantages should therefore come, it says, from widened opportunities. NRDC's 'commercial know-how and knowledge of industry's needs should also be brought to bear more directly on programmes of government industrial R and D, to involve the Government's research staffs more fully in the exploitation of their inventions.' (Trend also wanted to engulf

NRDC and thereby caused the most outspoken resistance.) Finally there is the AEA, which has already gone a long way towards industrial participation.

The Section called 'current problems' summarizes the arguments for a new organization, in part the same arguments as those used in the Trend Report, viz., the flexibility, independence, etc. of an organization outside the Civil Service. It faces also a serious defect in present arrangements in that the customer, so to speak, is always right. It is he who has to pay and knows what he wants, and so brings the problem to the supplier, i.e. the researcher, inventor and developer. Thus a contractual arrangement between research establishment and industry is to be encouraged. This argument can be contested in that an industrialist who can define his problem is on the way to its solution, but the Green Paper's point, that there must be good contact between the purchaser and the researcher, is valid.

The arguments about the AEA are not so convincing. For example, it says that the CEGB's need for research and development in nuclear power will expand as new nuclear systems are commissioned. True, but what about the CEGB's own research establishments, one of which (Berkeley) was built specifically for nuclear research and development? Is it so small that the CEGB will have to come to AEA establishments? Perhaps so, and the CEGB does do a little business with the AEA, but the Green Paper argument sounds rather specious. It also admits clearly that one of the arguments for keeping the reactor establishments together with the research establishments is that the former are running down sharply in staff numbers. Another argument is that the companies manufacturing nuclear power stations must be more closely associated with reactor research. This also may or may not be wholly true, for the companies have had experienced research departments of people concerned in nuclear-power generation from the days of Bradwell onwards. They have not, it is true, their own reactors in which to put test rigs, but even in this matter there is the fact that a great mass of experimental data is available from the AEA for any new system, and there is usually a prototype. So the argument here may be over-optimistic about the possible income for the new proposed organization. In the long view, when fast reactors come into regular use, maybe the Green Paper is right.

Certainly, if it should happen that Britain succeeded in selling power stations overseas, the companies would then need extra facilities in research and development. However, it should be noted that Britain has not sold a single nuclear power station overseas since 1957. After Calder Hall, when we were riding on the crest of the nuclear wave, Japan and Italy bought magnox stations. And these were all, regrettably. By now, if we were really out to sell, we should have sold the SGHWR overseas. Its capability for being built in any size without serious modification, its economy, its lack of need for specialized building techniques except for the nuclear centre; all these combine to make it a real competitor to American water-reactor stations, which have sold in great numbers round the world. Yet we have not sold one though we have an agreement of a sort with Greece, dependent in some way on our importation of Greek tobacco. (Will it be ground nuts in Nigeria?) It can be doubted whether we shall export an AGR, for it is competitive, certainly in Mark II, only in large sizes (1200 megawatts in Dungeness B), though this may change with more advanced types. However, nuclear power will certainly become a greater proportion of new power stations and eventually all new stations in Britain will be nuclear. So in the long run the arguments for the reactor establishments may be right.

Part III is the proposal itself for the British Research and Development Corporation (BRDC). The establishments to be included have already been named but it may be useful here to list them in tabular form with the number of professional staff and the gross annual cost of each to the Government (see table 1).

Each of these merits some discussion. The National Physical Laboratory is one of the three standards and measurement laboratories of the world, the best known of the others being the National Bureau of Standards in the U.S.A. The NPL has existed since 1900 and has built up a wide range of research activities and measurement techniques. It has had its variations in prestige and achievement over the years and is now riding high. It was at the NPL, for example, that Dr. Essen established his caesium frequency standard, which is now

**Table 1**

| | Professional staff numbers | Gross annual cost incl. capital expenditure (£m) |
|---|---|---|
| AEA | | |
| AERE, Harwell | 1200 | 16·4 |
| Culham Laboratory | 250 | 3·1 |
| Winfrith | 450 | 5·2 |
| Dounreay | 300 | 13·5 |
| Risley | 650 | 8·8 |
| Ancillary facilities elsewhere | 350 | 6·1 |
| National Physical Laboratory | 700 | 6·0 |
| National Engineering Laboratory | 350 | 3·1 |
| Warren Spring Laboratory | 200 | 1·3 |
| Hydraulics Research Station | 100 | 0·6 |
| Forest Products Research Laboratory | 100 | 0·4 |
| National Research Development Corporation | 100 | 5·2 |
| Total | 4750 | 69·7 |

Rounded figures for debate would be 5000 professional staff and a cost of £70 m.

*Note.* The cost for the NRDC is given as in the Green Paper, from which all the figures are taken, but does not appear in the latest annual report and accounts.

accepted as the international standard of time. The making of cheap diffraction gratings was pioneered here and so was the use of Moiré fringes in fine measurement. Thus one could go on but there is no need, for everyone knows the NPL. Its facilities include high-voltage laboratories and wind tunnels of every type much used in work on almost all the aircraft that have been built in Britain. (Modifications to the VC 10 were made as a result of work here.) It has established a unique place in the world of physics.

The National Engineering Laboratory was established at East Kilbride near Glasgow after World War II and has built up an enviable reputation. Hydrostatic transmission, air bearings, metal-forming techniques and so on are part of the work in which it is well known. Its expertise on pumps and fans is unequalled anywhere and has been called on for various overseas problems. It has been closely concerned with MinTech in the use of the Programmes Analysis Unit. On the numerical control of machine tools considerable work has been done. It devised the 2CL programme to replace the American APT.

Warren Spring Laboratory is at Stevenage. In the ten years or so of its existence it has earned over £½ m by consultancy and advice for the mining and metallurgical industries. Its main preoccupations are with chemical engineering and mineral technology. Many of its earnings have far outweighed the cost of the research. Processes for screening minerals and for extracting metals from poor ores have been invented in the laboratory.

The Hydraulics Research Station at Wallingford is unique. It is staffed almost entirely by civil engineers and its job is mainly to construct scale models, tidal where necessary, of harbours, bridges, dams, estuaries etc. and then find working data to improve flow, restrict silting, facilitate docking and so on. It also does research on such things as fluid flow as part of the foundation for its sponsored projects.

The Forest Products Research Laboratory at Princes Risborough serves the timber industry with research on problems of infestation, mechanized machining, the testing of timber and structures and joints and so on. It has, for example, helped mills in Britain to produce four times the wood pulp they did ten years ago.

The NRDC was established to facilitate the exploitation of British inventions by help in patenting and so on. It was loaned money at first by the Board of Trade and then was taken under the wing of MinTech. It has had its failures such as the steam engine fuelled by wood, constructed as a result of Sir Edward Bullard's committee that reported on solar energy

twenty or so years ago. (This committee showed that the most economic way of utilizing solar energy was to let it grow trees and then burn them to produce power.) But India, for whom the engine was intended, did not wish to pay even for the one built. NRDC has in fact written off some millions of pounds on failed projects over twenty-one years, a sign of the risk-taking involved in exploitation. It has had remarkable successes and has earned £2 m from licences on an antibiotic called cephalosporin discovered by Florey and others at Oxford. It put the Hovercraft on the map. It did the same for the Bacon fuel cell. It has become tough and selective about projects to support and has at last shown a small profit. In its departmental heads it has a body of experts on a wide range of technology and its exploitation.

There is no need to say much more than has already been said about the U.K. Atomic Energy Authority. Britain has made and tested its own nuclear weapons and has developed the first successful commercial-scale nuclear power stations. Some 12% of electrical power in Britain is produced now by nuclear fission, all based on AEA prototype stations, which followed years of research and development. The establishment at Winfrith has the High Temperature Gas-Cooled Reactor and the Steam Generating Heavy Water Reactor, the only ones in Europe. The AEA pioneered the use of natural uranium in power reactors. Its research is famous. Very recently it has turned towards non-nuclear work for industry. The Culham Laboratory is entirely concerned with fusion and plasmas and has a team of first-quality physicists and engineers. Large reactor groups are at Risley and Dounreay, where the prototype fast-reactor power station is being built, the first in the world.

All of these would be included in the new BRDC. Its aims and functions would be:

'(i) to encourage and support the development and application of innovation and technological improvement in industry for the benefit of the U.K. economy; and to carry out research and development for this purpose, both itself and in collaboration with industry and on repayment;

(ii) to carry out research programmes necessary in the public interest, including basic research, and other specific programmes of work required by Government departments and other public authorities;

(iii) to exploit where appropriate innovations resulting from Government financed programmes carried out by other agencies.'

The Green Paper goes on to forecast that Government work would provide the greater part of the Corporation's revenue for some years to come and that the biggest single element of the programme would be in reactor development. MinTech would appoint the Corporation Board and arrangements would be necessary for MinTech to determine or guide the broad lines of programmes. The BRDC would however be free to undertake on its own initiative work or investments on which it expected to recover its costs and would be able to engage in joint ventures with industry, sharing costs and risks. It would be required to arrange for most of its work to be undertaken on a contractual basis. Thus Government departments would pay for their own work done by the Corporation. Though obviously there would consequently be redeployment of money there would be no increase in total Government expenditure. The AEA would lose its advisory capacity for the Government and a new Atomic Energy Board with representatives from the nuclear industry companies and the generating boards would serve on it to advise on the size and shape of the reactor development programme. Earnings from contracts, joint ventures, royalty arrangements etc. would eventually make it possible to reduce the Government subsidy for industrial research.

The Corporation would be, the Green Paper believes, a powerful organization employing at the beginning nearly 5000 professional staff and costing about £70 m a year. (The Earl of Bessborough, in the House of Lords, gave a figure of 180 000 for total employees. Is this a misprint for 18 000 or his Lordship's mistake?) It is worth noting that the AEA part of the professional staff would be twice that of all the others combined and over four times the number of the largest single establishment, the NPL. It would be over thirty times as numerous as staff from the Hydraulics Research Station or the Forest Products Research

Laboratory. In short, the Corporation staff would be predominantly from the AEA. As one research director said: 'They will eat us alive'.

It is a sentiment shared by many of the establishments that would be taken over. Hardly one of them looks favourably on the proposed change. Each of course has its own particular case to present and is using all its skill in advocacy to safeguard its own position. Let us now consider tham all from the point of view of their contributions to, and connections with, industry, and of their incorporation into BRDC.

First consider the AEA. It has had considerable achievement in the non-nuclear field in the past three or so years and has done work for industry. The Green Paper says the AEA is making a quarter of a million pounds a year from industry. This official figure has been much quoted and it is wrong. Harwell alone, selling research and licences and services, has made in the financial year just ended over £½ m from non-nuclear R and D for private industry. Add to this the research and services in the nuclear field sold to the trading concern that will be British Nuclear Fuels. This comes to more than £600 000. Together the total industrial earnings for the year come to £1·2 m. It has also earned about £1 m credited from Government sources for non-nuclear work—desalination and carbon-fibre development. Indeed Harwell claims altogether £2·9 m earned 'outside the vote'. The range of work and services offered outside the nuclear field by the AEA is impressive. Its work on desalination by electrodialysis has helped William Boby to construct plants in Italy and Libya. There is pilot-plant operation of freezing and reverse osmosis. On multi-stage flash distillation, in collaboration with Weir Westgarth, its work is incorporated in commercial designs. It offers a materials-testing service. In ceramics, general analytical techniques (e.g. the first infrared temperature device that displays the temperature digitally) and so on it is creating good industrially-orientated development. Its computer time-sharing system is the biggest in Britain and it earns some income from educational and training courses even at post-graduate level. What there was of a feeling of superiority on the part of the scientists is surely gone and nobody can deny its energy in turning towards general technology. It has most to gain from belonging to the BRDC. Even from a least optimistic point of view a respite from the run-down would be achieved.

No serious person could look with equanimity on this run-down. Over 3000 people of high quality are involved. The nation needs this skilled manpower. There must be sympathy with the Minister's dilemma and appreciation of his attempts to resolve it by the creation of the BRDC. It is a human problem and a serious national one. It may be that the decision of the AEA to go for natural-uranium reactors was wrong. We might have gained exports with other systems. The much-boosted British lead in the production of electrical energy from nuclear fission will be lost inevitably as America catches up. It may be true also as Lord Blackett once suggested that there was an error after the war in including so many top scientists in Government establishments. We are left now with the problem—what to do with a surplus (1000? 2000? Who knows?) of experienced engineers and scientists, many of them first-class physicists, who will eventually be declared redundant unless something is done or unless they can inside the AEA earn many times what they are earning now from industry, impressive as the improvement has been.

The question can be asked however whether the BRDC is the correct way to salvage them. It could be, as one leader of industrial research, involved in the hard realities of commercial competition, has said, that we would be merely creating a self-perpetuating mechanism. He is quite harsh: sack the redundant ones, give them a good golden handshake to soften the possible reduction to lower salaries and prestige, and say good-bye and wish them well in getting posts elsewhere. There is some force in his argument, for Britain, if she is to have a growth rate worth having, cannot afford to carry a dead weight of superfluous men and women. It could be that this is exactly what the BRDC would be doing and when we examine the figures of the business that must be built up we may come to the same conclusion.

There is also the question of Culham. The Minister has agreed that basic research is necessary but why take over such a centre of excellence and put it in contact with the counting-house fingers? It costs little, about £3 m a year, and could be made a separate Fusion Centre

under MinTech. Dr. Adams a few years ago gave it 25 years before a practical scheme for power from fusion would be available, but if the technological problems should be solved at Culham it would become the newest prestige money-earner.

The non-AEA establishment with the biggest income from outside (and the smallest staff) is the Hydraulics Research Station. It makes about £500 000 a year, some 80% of its budget. Its work is unsurpassed anywhere in Europe. It already integrates with industry. It already lives on repayment. It suffers from no overlap of research projects with anyone else. It does not need any spur to make it profit-orientated. Why interfere with it administratively just for the sake of more organization?

The Forest Products Research Laboratory has the smallest annual income from outside, £30 000 at the most. This is because there is little money in the timber trade for more research. What advantage would there be for it or for the country in incorporating it into BRDC?

The Warren Spring Laboratory has very good connections with industry and covers many disciplines—physics, chemistry, mineralogy etc. Its income from industry is fairly small—£130 000 a year or thereabouts—but there is national interest in such matters as air pollution, in which it is expert. The country might gain from its inclusion in BRDC, but it is difficult to see what the Laboratory itself would gain.

The NPL would not benefit at all. It is a national institution, a large self-contained complex with duties not applicable to any other laboratory in Britain. It is unique in this country and has necessary liaison with similar laboratories in other countries. Its income from outside is £350 000 a year, a reasonable amount in view of the fact that it is not in existence specifically for repayment. It costs little to run for its size. It is moving continuously towards more effective man-hour productivity. Who would want to change it to be part of a large concern whose business is to be trading?

The NEL is difficult to evaluate in this context. Its income from outside is small compared with others. It of all the research establishments is more concerned with heavier engineering, which is what should be encouraged according to Lord Balogh (House of Lords 5th February). One can see that a BRDC would benefit from its inclusion, but if this were to interfere with the good connection built up with Scottish industry then the NEL would be seriously disturbed and so would Scottish industrialists.

In what form is the NRDC supposed to be included in BRDC? The Earl of Bessborough tried to elucidate this in the House of Lords. He quoted the Green Paper as saying 'assimilated' in one place and 'integration' in another. The reply from the Minister of State was that the intention was that NRDC should be 'absorbed'. BRDC might benefit from the fact that the NRDC when inside would be able to handle the licences at present dealt with by the AEA. If inclusion means more administration many NRDC men will react as Lord Black did in 1963. Is there someone inside the Government set-up that wants to get his hands on NRDC, just as there seemed to be in 1963? It is not easy to see that an NRDC as a department inside BRDC would do any better than it does now outside, except that wider scope has been hinted in the Green Paper.

Some general points about the proposed reorganization can be made. The first is the likelihood of income. The annual incomes of the constituent bodies at this moment are shown, rounded off, in Table II. All are exclusive of government credits or payments, and so MinTech orders are also excluded. The sums thus represent the incomes that the BRDC would wish to increase.

The total non-government annual income at present is thus £4·78 m. At a press conference the Minister said that he anticipated a 1 : 1 : 1 relationship between payments to BRDC by grants-in-aid, Government payments, and industrial payments. As the gross annual cost of all the component parts of the proposed BRDC is about £70 m then the expectation is that a third of this, or about £23 m, will be met by industry. What evidence is there to show that industry will step up its payments for research about five-fold (by 400%)? Even if good salesmanship could raise the industrial income by 10% per annum it would take 17 years to reach £23 m from £4·78.

**Table 2**

| | Thousands of £ |
|---|---|
| U.K. Atomic Energy Authority | 1200 |
| Hydraulics Research Station | 500 |
| National Physical Laboratory | 350 |
| National Engineering Laboratory | 180 |
| Warren Spring Laboratory | 130 |
| Forest Products Research Laboratory | 30 |
| National Research Development Corporation | 2390 |
| Total | 4780 |

Arithmetic of this sort suggests that the anticipated resurgence of the industrial use of research and development is unrealistic. Researchers are getting more and more profit-orientated, as the Green Paper has implied. Apart from salesmanship it is not they who are so much at fault, though some of course are. It is surely that industry is not convinced of the need for innovation based on research. Industrialists who do not have their own in-house research capabilities are showing only a moderate interest, if that. They could get the research if they wanted it from existing establishments and there is no need for a new organization such as the BRDC. We do need salesmen of research and promoters of innovation—troubleshooters of technological difficulties. A BRDC would do nothing to create these.

If the purpose of the BRDC on the other hand is to salvage the AEA, why not turn what there is of it into a corporation on its own? It could be the Atomic and General Research and Development Corporation (AGRDC) or A and G for short. It would then be able to reorganize itself and run its own affairs, though still dependent on some Government payment. Being made aware of cost-effectiveness it would deal with its own redundancies. It would give a new lease of life to the remaining AEA. It might even be able to emulate the big American contract-research organizations such as Battelle, Arthur D. Little and so on. It would be a challenging competitor to existing research organizations, which suffer more from industrial indifference than from inability to meet what demands there are. As it is, our own independent research firms do very small business and the biggest, IRD, just about reaches £1 m. My proposed A and G might fail I fear though it is worth considering.

There is another general point to be made. The Minister seems to think that reorganization is one of the cures for the present low usage of innovation. But organization, which is after all merely human topology, depends on people. We need to search for those exceptional men who can run research institutions and can establish the right sort of organization for creative scientists to do their best work; and a very subtle relationship it is. Reorganization in itself, beloved by the compulsive classifier, achieves very little.

If the Minister really wishes the image on the cover of the Green Paper—*A New Organization for the Seventies*—to be convincing to thoughtful and responsible citizens he could give more urgent thought to a couple of sentences of Part III (para. 35). It is the matter of pension. It sounds perhaps unrelated to great things like reorganization, but it is essential. The inability of scientists to move from university to industry or one firm to another or one corporation to another or industry to university etc. because of a fear about pension is a stultifying thing that deadens the maturing scientist's enthusiasm. In his twenties he may not care much. In his thirties and with a family he worries. In his forties he is scared to move. In his fifties he dare not. This is so fundamental that it outweighs any reorganization of research establishments. And there is one final point. Innovation should be made advantageous to industry. Awards and honours are not enough. It is the taxation system that needs close examination. Immediate solution of the problem of transferability without loss of pension and of the problem of taxation that takes no account of innovation would do more for the image than any Green Paper, which is part of an exhumed Trend Report brought up to date.

# Book reviews

***Gallium Arsenide, Proceedings of the 2nd International Symposium, Dallas, Texas, October 1968*** (Institute of Physics and the Physical Society, Conference Series No. 7). (London: Institute of Physics and the Physical Society, 1969.) Pp. vii+244. £6 5*s*.

This volume contains 35 papers presented at the above symposium, and represents a reasonable summary of work on this material in the two years since the first conference at Reading in 1966. The papers cover research in four areas: material preparation and assessment, devices based on optical effects, microwave devices and junction devices. Since 1966 there has been a change of emphasis away from microwave applications back towards the study of materials problems, with the realization that there is still insufficient control of processing techniques to enable good quality devices to be produced reliably and reproducibly. The main materials emphasis is still on epitaxial deposition, and one complete session of six papers was devoted exclusively to 'Liquid phase epitaxial growth', i.e. the growth of epitaxial films from GaAs solutions in Ga. This technique received only brief mention in 1966, but it has now been developed to produce films with higher carrier mobilities and lower carrier concentrations than films grown by vapour phase processes.

Although many of the papers covering materials aspects contain useful information on the processing conditions necessary to achieve optimum electrical properties, they are generally disappointing since a large number of them are more in the nature of reports prepared for Government sponsors than scientific papers. Thus there is a great deal of cataloguing of properties, such as carrier mobilities, concentrations and profiles, but little attempt is apparently being made to understand the basic processes involved. The papers on devices in many cases reflect the problems of fabrication technology. The progress towards solving these empirically is slow, and no major advances were reported.

The manner of presentation of the Proceedings has certain irritating aspects. It is inconvenient to have photomicrographs separated from the papers with no indication in the text that this has been done. It is inexcusable to have the photomicrographs from one paper (No. 9) in two separate places 63 pages apart. A number of unnecessary typographical errors also occur, beginning in the preface, where the maximum electron mobility measured at 77 °K is given as 13 500 $cm^2 v^{-1} s^{-1}$.

In spite of its shortcomings, however, the book contains material essential to workers in the field of GaAs technology.

B. A. JOYCE

***Thin Film Transistors—a new approach to microelectronics,*** by A. C. TICKLE. (Chichester: J. Wiley, 1969). Pp. ix+144. £4 4*s*.

This is the first book to be written exclusively about the thin-film transistor or 'Thin film insulated-gate field-effect transistor' to give it its full title. Since this is a relatively new technology, the amount of accumulated data is still limited and the author has been able to include details of most of the important features without producing a cumbersome volume.

The book covers both the experimental and theoretical aspects of the subject in sufficient breadth to enable the non-specialist to evaluate this new technology. Sections are included on the physical behaviour of semiconductors and insulators in thin-film form, particular

attention being paid to the influence of interface effects on possible conduction and modulation processes. The theoretical analysis of this type of transistor is treated in detail as are the various fabrication techniques. The author has used his considerable experimental knowledge to outline procedures suitable for analysing and optimizing device performance. These procedures should be of great assistance to the beginner in this field.

The book contains the usual sprinkling of typographical errors but they do not introduce any ambiguity. An irritating feature is the number of diagrams which have not been printed on the same page as the relevant text. By far the most important criticism however concerns the almost complete lack of reference to tellurium as a suitable elemental semiconductor replacement for the more complicated compounds, CdS and CdSe. This is a major omission and certainly detracts from the completeness of this text. However, this book will still provide the non-specialist with sufficient background information to enable him to gauge the usefulness of this form of microelectronics.

J. D. FILBY

***Applied Solid State Science, Vol. 1, Advances in Materials and Device Research,*** edited by R. WOLFE. (New York: Academic Press, 1969.) Pp. x+404. £7.

This is the first volume of a series of review articles, each "motivated by a class of devices ... newly invented or already widely used". This edition contains four articles; Junction Electroluminescence by P. J. Dean (pp. 140), Metal–Insulator–Semiconductor (MIS) Physics by A. Goetzberger and S. M. Sze (pp. 80), Ion Implantation in Semiconductors by J. W. Mayer and O. J. Marsh (pp. 100), and Electron Transport through Thin Insulating Films by S. R. Pollack and J. A. Seitchik (pp. 40). The articles are aimed at scientists starting research in the subjects discussed or working in closely related areas. Sections, however, should prove useful to workers well established in these areas, particularly to those concerned with device fabrication. The articles are clearly written and the subjects treated in a logical manner, making the texts ideally suited as a basis for further study.

The first article by Dr. Dean sets a high standard for the future and produces a very useful and up-to-date review. The introductory sections deal with the energy-band model as applicable to luminescence and lead to a discussion of the injection mechanisms at p–n junctions, heterojunctions, MIS contacts, and in graded band-gap materials. The author carefully works through the various types of recombination processes which have been identified to date using examples of spectra from the well studied materials GaP and CdS. Subsequent sections deal with semiconductor lasers and the article is rounded off with sections on 'state of the art' efficiencies, future prospects, and a brief survey of recent work (1968–69).

The non-mathematical treatment in the first article is contrasted by that on MIS physics in which various deductions are worked out in detail. Starting with the ideal MIS model, the influence of surface states is derived and the measurement techniques used to investigate their presence discussed and compared. Subsequent sections deal with the effects of charge in the insulator, metal work-function differences, crystal orientation, and the review is concluded with sections on the application of MIS diodes.

The article on Ion Implantation deals with the penetration and range of ions in semiconductors, the effects of bombardment and the subsequent annealing of damage and the re-location of the implanted ions. The electrical properties of the resultant material are discussed and the device applications of the technique reviewed.

The last article deals with the factors limiting the transport of electronic charge through insulators, considering first the theoretical aspects of carrier flow at contacts and in the bulk of the insulator, and then going on to compare the resultant theoretical predictions with the experimental results.

A subject index is included.

R. W. BRANDER

***Elementary Rheology***, by G. W. SCOTT-BLAIR. (London: Academic Press, 1969.) Pp. xi+158. £1 15*s*.

In about 150 pages, Dr. Scott-Blair takes us on a highly readable survey of the topics he has met in his forty years as a rheologist. He claims to be opening up the subject for technical people who are meeting rheology for the first time, and for students who have not been to one of the few colleges where rheology is taught. I believe that a third class will welcome it: the 'old hands', who will like to have a handy source book with references to some of the less well-known aspects of rheology.

As an elementary text on rheology, there is hardly an indication of the vast amount of work that has been done in setting up and applying constitutive equations for complicated materials. The emphasis is rather on the odd phenomenon, which, by careful experimentation, may be observed by a physicist, chemist, biologist, psychologist, in the course of primarily non-rheological investigations. However, there is a useful bibliography for those seeking deeper waters, and about a hundred references.

Anyone who thinks a technical subject should be presented in coldly objective terms may perhaps dislike the chatty style. The answer to that would be that the style is a product of an infectious enthusiasm that is all too rare in the technical literature.

It is a pity that the references to Bergson, Delboarf, Fechner and Weber are missing.

M F. CULPIN

***Advances in Materials Research, Vol. 3***, edited by H. HERMAN. (Sussex: John Wiley, 1969.) Pp. ix+417. £9 2*s*.

The new survey series, of which the volume under review is the third volume, appears to be taking over the role until recently fulfilled by Pergamon's *Progress in Materials Science*, itself the successor to a pioneering series: *Progress in Metal Physics*. The first volume of the new Wiley series contained short surveys of a number of common experimental techniques used in materials research; the second was concerned with experimental surveys of certain complex physical properties of solids, while the present volume turns more in the direction of theory. Of the five review articles in the book, only one, Segall's survey of fatigue-hardening in face-centred cubic metals, is mainly experimental in outlook. It is concerned with one restricted but vital aspect of fatigue, namely, the nature of the dislocation patterns induced by fatigue in metals of differing stacking-fault energy, complemented by an attempt to interpret these configurations. The author claims only modest insights, but these will be useful to workers in the field, especially because of the clear and straightforward way in which the evidence is set out.

Of the other articles, the first, by Mura, is concerned with the continuum elastic theory of dislocations, either singly or in groups. The treatment is highly mathematical, and ranges from well-known elementary results to involved relationships for dislocations in boundaries or near inclusions. There is no attempt to relate the treatment to 'real situations', such as yield-point, flow stress or fracture problems. The article appears to have been several years in the press; for instance, no reference is made to Nabarro's definitive text.

Wayman treats the crystallography of martensite transformations (mostly in steels) at considerable length. The principal established theoretical models are set out first, followed by an exhaustive survey of orientation relationship and habit planes, and a discussion of evidence concerning micro-deformation mechanisms in martensite plates. The article is clearly authoritative, as one would expect from the author, but suitable only for specialists.

Wallace's interesting survey of lattice dynamics and the stability of crystals is similarly of a level suitable for specialists, but repays skimming by others for the insight it gives into the physical principles which govern the mechanical stability of simple crystal structures and how these can be checked by computer model calculations. The instability of the simple cubic structure and the limitations on the stability of the body-centred cubic structure are clearly explained.

Finally, there is a short but intriguing essay by Mitoff on property calculations for heterogeneous (two-phase) systems. He does not attempt to survey the extensive literature but selects a few simple treatments only, explaining clearly why he rejects more complex models, and sets out the nature of the difficulties in predicting properties for particular two-phase systems. There is a brief section in which experimental measurements are compared with the results of the model calculation, but this part is inadequate; a useful critical survey would be even more valuable if it could, in a future volume, be complemented by a comprehensive survey of experimental results for two-phase composites, classified in terms of micro-morphology.

R. W. CAHN

***Magnetic Materials***, by R. S. TEBBLE and D. J. CRAIK. (Chichester: J. Wiley, 1969.) Pp. xiv+726. £12.

The book gives an authoritative, comprehensive and up-to-date survey of the magnetic properties of a wide range of materials. The technical exposition is concise and balanced, perhaps reflecting the combination of physicist and chemist present in the joint authorship, and the format and general layout adopted by the publishers are pleasing and such that the subject matter is easy to follow. The book surpasses in range and depth other recently published books dealing with magnetic materials and may well take over the role of Bozarth's *Ferromagnetism*, now in need of revision and modernization, which for many years has served as a reference book for research workers concerned with the behaviour of magnetic materials.

The book starts by giving the current theories in magnetism and follows by describing in turn the magnetic properties of the transition elements, the alloys and compounds of the transition elements, the rare-earth metals, the spinel ferrites, the rare-earth garnets, the orthoferrites and ilmenites, and the hexagonal ferrites. Each of the commonly encountered groups of magnetic materials, namely the permanent magnetic and soft magnetic alloys and ferrites, is then discussed at some length and the book concludes with an account of square-loop switching and memory elements. Each chapter is abundantly supported by a detailed list of references.

Perhaps the most controversial feature of the book is the choice of units adopted by the authors. They discuss the question of units in §1.2 and do in fact categorically state that the rationalized MKS system is used in the majority of standard text books in physics and is universally accepted in electrical engineering. Despite this, they say, "the greater part of the research literature in magnetism published to-day is in emu" and it is on this premise that the cgs system is adopted. Whilst sympathizing with the authors in their views I cannot help but feel that this is a retrograde course to follow. The MKS system is now accepted universally and no matter how incongruous some of its units appear in magnetic measurement it would seem that they should be used whenever possible to preserve uniformity. Nevertheless, despite this use of cgs units, I feel the book presents such a wealth of information on magnetic materials that it will be regarded as making a first-rate contribution to scientific literature and will prove of great value in all research centres.

F. J. WILKINS

***Fracture, Vol. 5***, by H. LIEBOWWITZ. (London: Academic Press, 1969.) Pp. xv+525. $21.25.

This volume, the fifth of a series of seven, is concerned with the fracture design of structures. Several authors contribute to chapters on, 'Fracture Arrest' (63 pp.), 'Fracture Design Practices for: Rotating Equipment' (93 pp.), 'Pressure Piping' (74 pp.), 'Ordnance Structures' (96 pp.) and 'Ship Structures' (92 pp.), 'Fracture Design Analysis for Air Flight Vehicles' (30 pp.) and 'Selection of Material and other Aspects of Design against Brittle

Fracture in Large Steel Structures' (52 pp.). Each chapter is a self-contained account of the 'state of the art' some twelve months ago; this does have the disadvantage that some material is repeated in the volume as a whole, for example the Pellini fracture analysis diagram appears in slightly different forms in three chapters.

The title 'Fracture Design' is somewhat misleading; a structure can readily be designed to fracture! What the authors consider are structures that will not fail under the designed loading or structures which will fail safe. It is generally recognized that practical engineering structures contain stress concentrations arising from design features and mechanical or metallurgical defects and that these defects can be extended during service as a result of creep and fatigue processes. Thus it is possible for failures to occur at stress levels below those permitted by normal design procedures. An important failure criterion is the critical stress at which a crack extends from a defect or stress concentration in an uncontrolled manner.

The critical stress for crack propagation depends on the size of the defect and the notch toughness of the material. In balancing the three factors for design purposes, the critical size of defect can be set, at the limit of detection of defects, or at the defect size which will cause a pressure vessel to leak before the crack propagates, or, as in aircraft structures, at the length of crack which is stopped by stiffeners. An alternative approach is to arrest a propagating crack before the structure fails completely as, for example, when up to a mile of gas piping fractures or a ship breaks in two. Bluhm, in the chapter on 'Fracture Arrest', indicates how special design features or the inclusion of tougher material within the structure can be used to limit fracture, and discusses the types of test for assessing crack-arresting conditions. After construction, many structures are proof tested above the design loading and this procedure is often beneficial to subsequent service life. Extensive service testing is used as a basis for the design of ordnance structures.

The various chapters include useful, well reproduced diagrams and illustrations and numerous references. The limitations of current theory and design procedures are recognized in each chapter and the authors indicate further research necessary to promote the design of more efficient safe structures. The publishers, editor and authors are to be congratulated on providing such a clear and up-to-date account of a subject which has grown rapidly in the last twenty years and continues to grow apace.

M. DEIGHTON

***Electrical, Magnetic and Visual Methods of Testing Materials***, by J. BLITZ, W. G. KING and D. G. ROGERS. (London: Butterworth, 1969.) Pp. ix + 202. £3 5*s*.

This book completes a three volume series covering conventional methods of non-destructive testing. X-ray and isotope methods and ultrasonic methods are covered by the two previous volumes so that this work is restricted to the three areas in the title.

The first chapter provides a very brief and elementary introduction to the basic properties of materials and outlines electrical and magnetic testing methods. There follows chapters devoted to eddy-current methods, magnetic particle testing, optical methods and penetrant methods. The detail and extent to which these techniques are discussed varies considerably. The magnetic particle method receives most attention and occupies two chapters, one on procedure and one on equipment. The section on eddy current contains some detail on equipment and is supplemented by a large appendix dealing solely with the operation of a specific commercial intrument (Forster's Magnatest Q). The chapter on optical methods is by far the shortest and takes a more fundamental line than the rest. This is because techniques such as interferometery and photo-elasticity are well understood, although their applications are limited. This chapter closes with a brief description of holography as a potentially useful testing method.

This book then is a comprehensive handbook for materials testers, or those interested in setting up production-line testing facilities. Many of the techniques described are qualitative

or comparative so that there is little in the way of 'theory' and the bulk of this well illustrated volume is taken up with detailed descriptions of testing procedures and the problems which arise in their application. There is no doubt about the practicality of the advice given, since Messrs. King and Rogers draw heavily on the experience they have gained in testing by magnetic and penetrant methods in the Rolls Royce Small Engines Division. One feels at times that the book degenerates into a collection of extracts from manufacturers' catalogues, but there is some virtue in having the information all together in one place.

D. MELVILLE

***Moiré Fringes in Strain Analysis***, by P. S. THEOCARIS, edited by J. SVED. (Oxford: Pergamon Press, 1969.) Pp. ix+426. £2 10*s*. (soft cover) and £3 (hard cover).

Professor Theocaris' comprehensive treatise is a must for anyone who wishes to apply moiré fringes for strain measurements and is not already skilled in their use. The book, divided into twelve chapters, covers the theoretical and technical aspects of the subject. After a brief introductory chapter, the author deals at length with the moiré patterns formed by line gratings in chapter 2 and by circular, radial and zone gratings in chapter 3. These chapters contain full mathematical treatments of the fringe patterns formed, both from a general standpoint and for particular cases.

The next six chapters deal with the technological applications to strain measurement—firstly the use of line gratings for two-dimensional strain measurement, then the use of remote gratings, superposing an image of one grating on to another after passing the light through a transparent specimen subjected to plane stress, so that the pattern formed is due to the distortion of the specimen surfaces by the Poisson's ratio effect. Methods for studying surface topology, deformations under dynamic conditions and strain measurements on curved surfaces and the use of the moiré phenomena in extensometers are discussed. Specific examples of the use of each are given, several being Professor Theocaris' own work, and all are well illustrated with graphs and photographs.

The next two chapters deal with the equipment, firstly the precision of gratings and the influence of defects, and then information on the techniques for the production of gratings. The final chapter discusses the advantages and disadvantages of moiré methods for strain analysis, listing in particular that they determine strain distributions directly from purely geometrical relationships without dependence on measurements of other physical properties such as electrical resistance or birefringence, that they can be used at high temperatures with suitable grids etched on to the specimen, that small elastic deformations and large plastic deformations can be determined with almost equal ease and accuracy.

The book is easy to read and the quality of the production is high.

K. J. PASCOE

***Advances in Microwaves, Vol. 4***, edited by L. YOUNG. (New York: Academic Press, 1969.) Pp. xiii+384. £8 3*s*. 4*d*.

This volume differs from its predecessors in the series in that it has only three chapters, of which the first occupies over half the book. With chapters of this length one would expect a thorough treatment of the subject matter, and in this one is not disappointed.

The first chapter deals with devices for introducing phase shifts and time delays in microwave transmission systems with particular application to electronically steered aerial beam patterns, and discusses the uses of ferrites and diodes to achieve the effects needed. This is done in a way which keeps physics in mind, but nonetheless provides sufficient engineering data to enable the would-be user to design his own devices. The discussion of ferrites is particularly valuable, but the treatment of the diode methods could with advantage have been backed up with more design information. Time-delay elements are

reduced to lengths of stable transmission line and switches: There is no mention of the acoustic delay line as it has no place in aerial systems.

The second chapter deals with millimetre waveguides and the Japanese railways, a somewhat esoteric combination which however contains a very useful discussion of several novel types of millimetre-wave transmission line, particularly various configurations of dielectric rod and tube lines. This is followed by a discussion of leaky waveguides and beam waveguides, for the ultimate interest lies in the use of microwaves for obstacle warning and train communications on high-speed railways. The chapter ends by describing the results obtained from various experimental installations.

The final chapter is mainly concerned with mode conversion in tapered transitions between waveguides of differing size or shape, and is heavily mathematical. The technique used is to reduce Maxwell's equations to a set of generalized telegraphists equations, thence to a Volterra integral equation which may be solved if the waveguide transition is sufficiently gradual. This chapter is aimed at those to whom unwanted modes at the $-60$ dB level cause concern, rather than those of us who are not sure that they could measure power at this level, let alone worry about it.

In its general presentation the book follows the high standard of the previous volumes, and is lavishly illustrated with diagrams and photographs. The ample reference lists extend to mid-1967.

D. J. H. WORT

***Handbook of Vacuum Physics, Vol. 2, Physical Electronics* (*Parts 4, 5 and 6*)**, edited by A. H. BECK. (Oxford: Pergamon Press, 1968.) Pp. vi+200. £3 3*s*.

The book under review is one of a series which covers a far wider field than one would expect from the general title. This particular half volume has three chapters. The first, by E. A. Ash, is devoted to electron optics, the second by A. H. Beck, to heavy current electron beams and the third, also by A. H. Beck, to the theory and design principles of gridded valves.

The chapter on electron optics starts with a theoretical treatment of electron lenses and then gives the theoretically predicted characteristics of several idealized electron lenses. Methods for determining field distributions and for calculating electron trajectories through practical lenses are then described. After a treatment of aberrations there is a section on cathode imaging and electron mirrors. The final section contains a useful treatment of beam deflexion, post-deflexion acceleration and deflexion aberrations.

The second chapter is mainly devoted to the design of electron guns, and includes several practical examples. Various methods of confining and focusing high-current beams are described, and there are useful sections on annular and strip beams.

The third chapter is a short and mainly theoretical treatment of gridded valves.

The book is well printed with a semi-stiff cover. The diagrams are clear and each chapter has a reasonably comprehensive list of references. Unfortunately there is no subject index and this would limit the usefulness of the book for reference purposes. It is, nevertheless, a good, concise, introduction to electron optics and electron beam dynamics.

P. R. STUART

# Forthcoming courses and conferences

## INTEGRATED CIRCUIT TECHNOLOGY 1970

**14–25 September 1970 at Enfield College of Technology**

A short full-time course suitable for the graduate-level engineer who requires a reasonably detailed background treatment of integrated circuit design, fabrication and testing. Practical laboratory work in the College Microelectronics Laboratories will be combined with lectures. Fee is £50 (excluding accommodation).

Applications to the Academic Registrar, Enfield College of Technology, Queensway, Enfield, Middlesex.

## X-RAY TECHNIQUES IN THE INDUSTRIAL LABORATORY

**22–26 June 1970 Imperial College, London**

Organized by the McCrone Research Institute. This course will cover the whole gamut of X-ray techniques.

For further information apply to Course Registrar, McCrone Research Institute, 2 McCrone Mews, Belsize Lane, London NW3.

## SOLID STATE DEVICES

**15–18 September 1970 at the University of Exeter**

Organized by IPPS in collaboration with the IEE, IERE and the U.K. and Republic of Ireland section of IEEE. This is the fourth conference in the series, and as previously is aimed at presenting new work in applied solid state physics, device production and characterization together with associated technology.

Further details and application forms will be available in May–June 1970 from the Meetings Officer, IPPS, 47 Belgrave Square, London SW1.

## GALLIUM ARSENIDE

**5–7 October 1970 at Aachen, Germany**

Sponsored by IPPS and the USAF Avionics Laboratory. This is the third conference in this international series. The conference will cover the area of material and device technology of gallium arsenide, gallium phosphide and related compounds.

Further details are available from the 1970 International Symposium on Gallium Arsenide and Related Compounds, Technische Hochschule, 5100 Aachen, Templergraben 55, Germany.

## PHYSICS OF NON-CRYSTALLINE SOLIDS

**14–18 September 1970 at the University of Sheffield**

This conference aims to study the structure of inorganic and polymeric glasses, transformation phenomena and viscosity–temperature relationships for supercooled liquids and glasses; and the relaxation effects and mechanical properties of glassy materials.

Further information from Bryan Ellis, Department of Glass Technology, University of Sheffield, Northumberland Road, Sheffield S10 2TZ.

## CRYSTAL GROWTH

**23–25 September 1970 Zurich, Switzerland**

Organized by the Crystal Growth Section of the Swiss Crystallographic Society in collaboration with the Comite International de la Croissance Cristalline. The meeting will include a series of invited papers on major topics by leading authorities together with a programme of short (20 min) contributed papers on bulk growth of crystals from the vapour phase, epitactic growth of layers and whisker growth.

Chairman of the organizing committee is Dr E. Kaldis, ETH, Laboratorium fur Festkorperphysik, Gloriastr 35 CH-8006 Zurich, Switzerland.

## COMPUTATIONAL PHYSICS

**7–9 September 1970 Imperial College, London**

Organized by the IPPS's Computational Physics Group. This is the second annual conference and it hopes to cover the following topics: experimental high energy physics, meteorology, astrophysics, nuclear, atomic and molecular physics, fluids, numerical methods, space science, oceanography, crystallography and solids.

Further details and application forms will be available from June 1970 from The Meetings Officer, The Institute of Physics and The Physical Society, 47 Belgrave Square, London SW1.

## ULTRASONICS FOR INDUSTRY 1970

**20–21 October 1970 London**

Sponsored by the journal 'Ultrasonics'. The main topics of the Conference will be high power ultrasonics (macrosonics), non-destructive testing by ultrasonics, medical and biological ultrasonics, ultrasound in communication and control, underwater acoustics and microwave ultrasonics. An exhibition will be held at the same time as the conference.

Further details from Ultrasonics for Industry 1970, 32 High Street, Guildford, Surrey, England.

# Sources of Reviews

**2–8** ESRO and its programme by H. Bondi and J. Ortner was published previously in *J. Phys. D: Appl. Phys.*, **3**, 1–7.

**9–26** Molecular engineering in the design of microwave ferrimagnetic garnets by A. S. Hudson was published previously in *J. Phys. D: Appl. Phys.*, **3**, 251–68.

**27–46** Recent developments in conventional soft magnetic materials by J. E. Thompson was published previously in *J. Phys. D: Appl. Phys.*, **2**, 933–52.

A. VECHT and N. J. WERRING

Division of Materials and Molecular Science, Woolwich Polytechnic, London

# Direct current electroluminescence in ZnS

**Abstract.** This review deals mainly with direct current electroluminescence (dcel) in manganese and copper activated ZnS. The preparative techniques and dcel observations on powders, thin films and single crystals are described. The basic mechanisms are outlined and possible future developments are discussed.

## 1. Introduction

Electroluminescence in solids has been the subject of extensive study since the observation of light emission from silicon carbide by Lossev (1923) and from zinc sulphide by Destriau (1936). In the case of silicon carbide, Lossev observed emission from crystals under forward bias and the effect was explained many years later in terms of injection at p–n junctions (Lehovec *et al.* 1951). Destriau achieved electroluminescence only under alternating current excitation from zinc sulphide powders embedded in an insulating medium, and the effect is generally ascribed to the recombination of excess carriers generated by impact ionization in high-field regions.

Most of the effort on zinc sulphide has been devoted to alternating current excitation, and direct current electroluminescence (dcel) has not been widely observed. In the last few years, books and reviews have been published surveying electroluminescence (Piper and Williams 1958, Curie 1960, Henisch 1962, 1964, Ivey 1963, Fischer *et al.* 1965, Fischer 1966, Cusano 1967, Morehead 1967, Thornton, P. R., 1967). Most of these pay considerable attention to zinc sulphide, but the discussion of dcel is limited. In this review, we will concentrate on dcel in all forms of zinc sulphide, i.e. powder layers, thin films and single crystals.

## 2. Preparative techniques

The performance of electroluminescent phosphors is governed by the preparative techniques employed. Unfortunately, the details of the actual methods used in specific phosphors are not often reported. Usually, unspecified raw materials labelled 'luminescent grade' are treated to remove volatiles and, subsequently, the required impurities, such as manganese and copper, are introduced. The difficulty in correlating various techniques and results is further augmented by the virtually complete absence of any quantitative data on the extent of the deviation from stoichiometry in the materials studied.

### 2.1. *Powders*

The selection of possible methods for the preparation of activated ZnS powders has grown considerably in the last decade. As reviewed by Jackson and Vecht (1969), these methods range from the conventional precipitation of activated (or unactivated) ZnS using hydrogen sulphide or ammonium sulphide (Leverenz 1950), to homogeneous precipitation using organic precipitants such as thiourea. Other methods include solid-state organic reactions and the decomposition of organometallic compounds such as zinc diethyldithiocarbamate.

The copper, halogen and manganese impurities, as well as the variation in stoichiometry, can be introduced during any one of the following stages:

(i) During the precipitation or synthesis of the ZnS.

(ii) By 'slurrying' the raw material in solutions of the required impurities, followed by a firing stage.

(iii) By firing the doped (or undoped) raw material in the appropriate atmosphere.

As will be described below (§3), the importance of a high surface concentration of copper (or copper sulphide) to give good d.c. electroluminescence has been generally recognized. The coating of the phosphor with copper to ensure a copper-rich phase at the surface, has been shown to yield bright d.c. phosphors (Vecht *et al.* 1969).

### 2.2. *Thin films*

Three principal methods for the preparation of useful dcel thin films have been used. These are:

(i) Direct evaporation of activated ZnS (Goldberg and Nickerson 1963, Golovkina *et al.* 1965).

(ii) Chemical reaction of zinc and activator halides with sulphur-containing compounds (Studer and Cusano 1955).

(iii) Embedding and firing the thin film in a powder phosphor containing the required activators (Thornton 1963, Vecht and Ely 1964).

Method (iii) tends to be the most reproducible and some results have been reviewed by Vecht (1966).

The copper coating technique has been applied to thin films by Cusano (1962). Manganese-activated films were made d.c. electroluminescent by treating the zinc sulphide surface in a dilute aqueous solution of a copper salt. Subsequent washing in potassium cyanide removed the copper-rich phase and eliminated the dcel emission.

Recrystallization of thin films of II–VI compounds has been studied extensively by many workers (e.g. Gilles and Van Cakenberghe 1960, Dresner and Shallcross 1963, Te Velde 1964, Vecht 1966). There appears to be some correlation between recrystallization and electroluminescence (Vecht 1969) but this is not yet fully understood.

### 2.3. *Single crystals*

Single crystals of ZnS have been prepared from the vapour or from the melt. Growth from the vapour phase is usually employed because of the high melting point (1830 °c). Piper (1952) obtained single crystals from ZnS vapour in a sealed tube placed in a temperature gradient, whereas Greene *et al.* (1958) used atmospheres of argon, hydrogen sulphide, hydrogen or nitrogen. Kremheller (1955) transported the ZnS vapour in a gas flow of hydrogen sulphide and hydrogen or, alternatively, pure helium. Growth by controlled cooling of melted ZnS has been accomplished by Addamiano and Aven (1960) at pressures as low as 7 atmospheres under argon, but their optimum pressure was 150 lb $in^{-2}$.

All crystals have tended to lack reproducibility owing to the high density of defects and poor control of stoichiometry. The effects of non-stoichiometry on the electroluminescent characteristics of ZnS crystals have been high-lighted recently by Biter *et al.* (1969).

## 3. DCEL observations

### 3.1. *Powders*

Powder electroluminescent layers are, in general, insulating and the current flow is mainly capacitive. For this reason, steady light emission will only occur when alternating voltages are applied. The presence of an insulating binder plays a part in this, since it prevents powder particles (which may be conducting) from being in contact with the electrodes. Thus, one of the first observations of dcel in powders was made by Zalm *et al.* (1954) on a binderless layer. Light output was obtained only at high voltages (> 100 v on a 20 $\mu$m layer), and originated from a small number of spots. It is interesting to note that these workers recognized the need for a copper-rich surface layer to obtain electroluminescence. Since copper can only be incorporated in the ZnS lattice to a limited extent (see, for instance, Aven and Parodi 1960), excess copper will be precipitated as copper sulphide at the surface. The presence of this layer is almost universally accepted, and it is common to

most proposed mechanisms of dcel, although its role in the process can be interpreted in different ways.

The insulating nature of powder layers causes the build-up of space charge when free carriers are injected. The resulting effect is that an internal polarization field will be set up in opposition to the applied field. Any emission which might appear on the initial application of a d.c. voltage will soon fall. Further emission may be obtained when the d.c. voltage is removed; in this case, luminescence is induced by the polarization field. Thus, excitation by d.c. pulses can give rise to electroluminescence from insulating layers and many early experiments were conducted in this manner, e.g. see Butler and Waymouth (1955).

Nicoll and Kazan (1955) observed d.c. emission when high fields were applied to phosphors in cathode ray tubes. At about the same time, Bowtell and Bate (1956) obtained d.c. emission from compacted layers of phosphor compressed between two conducting surfaces. The phosphors used included zinc sulphide, and for the production of light, the phosphor particles were coated with a layer of an ionizable salt (usually potassium silicate). An electrolytic action was observed under initial a.c. (50 Hz) or d.c. voltage application, in which the build-up of light emission was accompanied by a fall in current. The formation of a barrier, conducive to the excitation of luminescence, at the electrode surface was proposed, and it was suggested that the phosphors used by Nicoll and Kazan (1955) may have been contaminated with potassium silicate or other salts during the screen preparation.

Zalm (1956) maintained that dcel in powders was a surface effect and assumed that an exhaustion barrier of the Mott–Schottky type existed at the interface between zinc sulphide and the conducting copper-rich phase. In his model, excitation occurs at the cathode and, when the electric field is applied, the barrier broadens owing to field ionization of donor levels. A similar mechanism was proposed by Piper and Williams (1955) to account for their observed a.c. behaviour in ZnS single crystals. Zalm derived an expression for the dependence of light output on voltage of the form $\exp\{-\text{const.}/(\text{volts})^{1/2}\}$. Both the supply of electrons by tunnelling through the barrier and the probability of excitation have the same voltage dependence. This relationship was experimentally verified by Zalm *et al.* (1965) for a.c. and d.c. excitation. The only d.c. phosphor showing this law over three orders of brightness was, in fact, ZnS–$Li_2S$–Mn. Similar voltage dependence was reported for a.c. excitation of single crystals by Alfrey and Taylor (1955). Zalm (1956) attempted to distinguish between 'fundamental' and 'characteristic' activators by means of the luminescent response to d.c. voltage pulses. The 'fundamental' activators (such as copper and silver) give rise to energy levels in the forbidden energy band and are capable of ionization. Owing to the presence of the necessary high fields, free electrons produced by ionization will be carried away from the excited region before they can recombine. Thus, when the pulse is on very little emission is observed. On the removal of the pulse, electrons can now return to the ionized region and give rise to a light flash. 'Characteristic' activators (such as manganese and rare earths) on the other hand, luminesce by means of electronic transitions within closed shells and ionization of the centres does not occur. Maximum emission then coincides with the maximum voltage, apart from a small phase shift due to the lifetime of the activator in the excited state.

Thornton (1958), better known for his results on thin films, observed dcel in conventional powder-plastic dispersions with high phosphor-plastic weight ratios. Although the d.c. emission was usually much less than with a.c. excitation, the addition of a d.c. voltage and an a.c. voltage resulted in a light level considerably greater than the sum of the separate emissions (Thornton 1959 a). An increase in the a.c. light output of up to 250 times was recorded by the addition of a d.c. voltage. From a study of the brightness waveforms, the major enhancement process was shown to be d.c. ionization of luminescent centres. The main light peak occurred on the half cycle in which the d.c. and a.c. fields were in opposition, since free electrons would be driven back to the ionized centres by the a.c. field. Recombination then takes place in a low-field region, while d.c. emission involving ionization occurs in

regions experiencing a high electric field. In a later paper, however, Thornton (1961 b) proposed a minority carrier injection mechanism to account for the observations of emission at less than band-gap voltages on both thin films and powders, and the similarities between the current–voltage characteristics of the phosphors and those of forward-biased semiconductor diodes. In Thornton's view, the difficulty with powder layers is to establish large area electrodes in good contact with the particles, and also particle-to-particle contact to enable current flow. No absolute measurements of dcel emission were published by Thornton on powder layers.

Kazankin *et al.* (1959) observed dcel in powder layers of ZnS : Cu, Mn in castor oil. Persistent luminescence was only observed when manganese was present and when sufficient copper was added to form a second phase of copper sulphide. The function of the copper sulphide was to increase the conductivity of the layers. Removal of the copper sulphide by washing in potassium cyanide solution caused a considerable reduction in the current and the layers now showed very little dcel. The operating life of these layers was rather short, deteriorating to half brightness in the order of minutes. Similar characteristics were reported by Favorin and Kozina (1961) on ZnS : Cu, Mn powder phosphors in a solid dielectric resin.

Various workers attempted to elucidate the mechanisms involved in the electroluminescence of ZnS powders by using d.c. voltage pulse excitation. Zallen *et al.* (1960) measured the emitted light pulses on the application and removal of the voltage pulse, and studied the effect of the pulse width and repetition interval at different temperatures. They confirmed that for yellow-emitting ZnS : Cu, Mn phosphors, bound states of the manganese centre are involved in the electroluminescent process and ionization does not occur. The build-up of a polarization field was indicated. A proportion of the polarization charge consisted of electrons in deep traps, and slow thermal release from these levels gave rise to the observed decay of the light pulses. Thermal activation energies between 0·3 and 0·8 ev were found and the traps were attributed to the host crystal rather than to the luminescent centres. Vereshchagin (1962) performed similar experiments on ZnS powders containing copper as the main activator. In addition to composite layers, measurements were carried out on an individual 70 $\mu$m phosphor particle. Emission was observed at the anode end of the particle on the application of the voltage pulse, and at the cathode on its removal. The proposed model assumes the existence of an internal p-type region surrounded by n-type material, and the excitation occurs when the surface p–n junctions are reverse-biased. The applied field excites the anode junction and a polarization field, built-up during the time the pulse is on, gives rise to emission at the cathode when the applied field is removed. By the application of a large voltage, Fridkin (1961) achieved polarization of ZnS; Cu, Cl powder layers which persisted for a long period of time. The internal field, caused by a non-uniform distribution of holes and trapped electrons, remained indefinitely owing to the negligible probability of thermal release of the electrons from deep traps. Subsequent voltage application, in the opposite direction to the original applied voltage, caused light emission due to the recombination of electrons freed from the deep traps by tunnelling. Blank and Yatlinko (1964) also induced polarization and observed emission with field reversal. The maximum brightness observed was about 10 nit ($\simeq$3 ft lamberts) and persisted for only a few tenths of a second.

Quantitative brightness and efficiency data on dcel powder layers on ZnS were first reported by Vecht and Ellis (1966). At room temperature a luminance of 10 ft lamberts was recorded in the ZnS : Cu, Mn system and a quantum efficiency of 20%, corresponding to 0·1% power efficiency was obtained. The firing atmosphere had a marked effect on the efficiency; the nitrogen-fired phosphors were an order more efficient than air-fired ones. The air-fired layers, however, were capable of operation at higher brightness levels before electrical breakdown occurred. The amount of copper incorporated also determines the brightness and efficiency characteristics (Vecht and Ellis 1968). With copper concentrations of about 0·1–0·2% maximum brightness was achieved, whereas the most efficient layers were obtained with copper concentrations in the region of 0·05%. Control of the materials

preparation has led to the realization of higher efficiencies and higher brightness levels (Vecht *et al.* 1968, 1969). High purity starting material and careful monitoring of the level and nature of the manganese and copper impurity additions were required to obtain reproducible dcel properties. Two types of phosphor have been prepared: the first by a slurrying technique, in which the activators are added to pre-fired ZnS powder (post-activation), and the second by the addition of the activators during the precipitation of the ZnS (simultaneous activation).

By optimizing the manganese, copper and chloride concentrations in post-activated phosphors, efficiencies as high as 170% quantum (corresponding to 0·33% power) have been achieved at room temperature at a luminance of 0·5 ft lambert. The variations of the impurity concentrations were investigated independently (Vecht *et al.* 1969). The observed rapid loss in efficiency owing to the presence of chloride in excess of 70 p.p.m. was significant. Temperature experiments indicated maximum efficiency and brightness at room temperature; operation at lower temperatures drastically reduced the performance. At constant voltage, the luminance dropped five orders from 300–120 °K and the quantum efficiency fell to $10^{-2}$%.

Layers showing very high brightness values were obtained by copper coating a simultaneously activated phosphor. A luminance of 1000 ft lamberts was recorded at an efficiency of 5% quantum and 0·05% power. Panels made from this phosphor type require much lower voltages (200 v for the values quoted above). Reduction of the layer thickness, from the normal 50 $\mu$m to 10 $\mu$m yielded an output of 10 ft lamberts at 30 v. The copper-coated phosphors exhibit a 'forming process', in which the initial high current falls during the first minute of voltage application if sufficient power is dissipated in the layer. The fall in current is accompanied by the onset of dcel emission. This process is believed to involve the local diffusion of some copper species. These highly conducting layers can be operated in either through, or gap, cell modes. Preliminary deterioration experiments were quite promising; a half-life of 900 hours from 10 ft lamberts was observed.

### 3.2. *Thin films*

Investigations on the dcel properties of ZnS thin films have, in general, proved more fruitful than studies on powder layers. Geometrically thin films can more readily be subjected to high uniform fields, since they can be prepared with uniform thickness down to fractions of a micrometre. The absence of dielectric binder materials enables one to make direct electrode contact to the phosphor. As with powder layers, insulating thin films do not exhibit good dcel characteristics. For example, Halstead and Koller (1954) observed no d.c. emission below breakdown, except for a transient flash on the application and removal of d.c. fields, on ZnS : Mn thin films having a resistivity of the order of $10^{13}$ $\Omega$ cm. However, under a.c. excitation, 'in-phase' electroluminescence was observed. In common with powders, the presence of copper, usually in excess of the solubility limit in ZnS, again seems to be essential for the appearance of light emission under d.c. excitation.

Thornton studied the dcel characteristics of ZnS thin films and made a considerable contribution to this field of research. He found that the incorporation of manganese, in addition to copper and chlorine, gave better d.c. results than with copper and chlorine alone (Thornton 1959 b). The light output was dominated by the characteristic yellow manganese emission. The observation of dcel at very low voltages was of great significance. After achieving a.c. emission at 1·5 v (r.m.s.) on a ZnS : Cu, Cl film (Thornton 1959 c), dcel at 2 v was obtained on a ZnS : Mn, Cu, Cl film (Thornton 1961 a). The outputs at these voltages were very low ($< 10^{-9}$ ft lamberts) and no indication of a threshold voltage was apparent. The appearance of light output at less than band-gap voltages, and the similarity between the current–voltage characteristics and those of semiconductor diodes led to the formulation of a mechanism involving minority carrier injection (Thornton 1961 b). The difficulties in attributing this model to ZnS are discussed in §4.2. Patek (1961) suggested that the impact ionization mechanism could still account for Thornton's low-voltage results if the influence of the thermal velocity of electrons was taken into account. In his main paper on dcel thin films, Thornton (1962) reported outputs of 600 ft lamberts at an applied

voltage of 85 v. A power efficiency of 0·05 lm $w^{-1}$ ($\simeq 10^{-4}$ w/w) at 10 ft lamberts was recorded. The emission of light uniformly throughout the film volume was implied by the observed increase in output with film thickness at constant current. The temperature dependence of current, voltage and light emission was found to be rather complicated. As the operating temperature was reduced from 300 to 100 °K, the emission increased from three to ten times, at constant current. At constant voltage, however, the same temperature change caused the emission to fall by three to ten times and the current to fall by fifty times, giving an order of magnitude improvement in efficiency. The maintenance of dcel at constant voltage was no more than a few hours to half of initial brightness. Thornton (1963) managed to reduce the voltage required for high light emission on films activated by his embedding technique. Brightnesses of 100 ft lamberts were achieved at about 35 v and, in one case, a one inch square area gave 50 ft lamberts using a 22½ v battery.

Cusano (1962) observed strong dcel in manganese activated thin films. The films, prepared by a vapour reaction method, were treated in a dilute aqueous solution of a copper salt. This converted a thin outer phosphor region into a copper-sulphide phase, which was found to be primarily $Cu_2S$ (also some $Cu_9S_5$) and exhibited p-type conduction with a resistivity as low as 0·2 Ω cm in some cases. Over 100 ft lamberts were achieved with d.c. excitation (compared with 1100 ft lamberts 'in phase' at 6 kHz), and the maximum power efficiency approached $2 \times 10^{-2}$%. Electroluminescence was only achieved when the metal contact to the $Cu_2S$ was negative, i.e. when the $Cu_2S$/ZnS junction was reverse-biased. A mechanism of majority carrier (electron) injection by tunnelling followed by collision excitation of the manganese centres was proposed. Films containing no manganese were only weakly d.c. electroluminescent, and power efficiencies of only $5 \times 10^{-4}$% were recorded.

Goldberg and Nickerson (1963) prepared thin films of ZnS : Mn, Cu, Cl showing dcel at low voltages. A double emission band (with peaks at 5700 Å and 6050 Å) was observed, and a brightness of 5 ft lamberts at 12 v was achieved. Using a refrigerated photomultiplier tube, light emission was detected down to 2 v. The power efficiency was independent of brightness level and ranged from $10^{-3}$ to $10^{-2}$ lm $w^{-1}$ (i.e. about $2 \times 10^{-3}$%). The life of the films to half initial brightness was some tens of hours. Using a gap-cell configuration, light emission was located at the anode with all the metals tested, i.e. Cu, Ag, Mg, Al and In. It was assumed that the emission from a sandwich cell also occurred at the anode, but only Al and Mg contacts were successful. With these metals, electrical shorting paths produced during the metallizing process burnt out, and an insulating oxide film formed during voltage operation which acted as a blocking layer. A forming process is described which is essentially the same as more recently found by Vecht *et al.* (1969) on powders. Although the Goldberg and Nickerson paper states that all electroluminescent films exhibited this phenomenon, the description is confined to sandwich cells. The proposed mechanism of the forming process involves the anodic oxidation of the Al electrode. If forming was also found in gap cells with other metals, e.g., Cu and Ag which oxidize at much higher temperatures than Al, then this mechanism may not be entirely relevant. Since a high concentration of copper was required for dcel, a process involving the diffusion of some copper species (as proposed by Vecht *et al.* 1969) may take place. Furthermore, during the preparation of the films, a copper layer was deposited on to the ZnS film and the film was then heat treated (Nickerson and Goldberg 1963). This would lead to the formation of a copper sulphide ($Cu_xS$) phase. The $Cu_xS$ phase would surround each crystallite, the size of which was reported to be approximately 1 μm (Nickerson and Goldberg 1963). Although Goldberg and Nickerson observed emission at the anode region of the gap cell, it is possible that electroluminescence takes place at the cathode end of a ZnS crystallite which is adjacent to the external anode. Some evidence of an electrolytic deposit of copper sulphide at the anode of a gap cell with copper electrodes was found. The proposed dcel mechanism based on the appearance of light at the anode, involves hole injection from $Cu_xS$ into ZnS and electron injection from the tin oxide conducting glass. Recombination occurs via the copper centres and resonance energy transfer leads to the manganese emission.

The function of a copper-rich layer in promoting dcel was demonstrated by Inari and Takeda (1964). A ZnS : Mn phosphor was evaporated on to a conducting glass substrate and during the last few minutes, copper metal was simultaneously evaporated. A metal electrode of Au or Al was then deposited. Uniform manganese emission was only obtained when the metal was biased negatively. Regions of p-type copper sulphide existing near the metal electrode would thus be reverse biased with respect to the ZnS. In common with Cusano (1962), tunnelling followed by impact excitation of the manganese centres was proposed.

Golovkina *et al.* (1965) achieved dcel in ZnS films with the metal electrode at a positive or a negative potential. The configuration employed, however, makes analysis rather difficult. They evaporated ZnS : Mn, Cu, Cl phosphor in three layers; the central layer having an additional strong copper doping. A thin insulating silicon monoxide layer was evaporated before the aluminium electrode. The films were characterized into two types; one giving dcel with the Al contact positive, and the other giving dcel with either polarity. The preparative conditions for the two types appear to differ solely in the thickness of the phosphor layers without excess copper and in the thickness of the silicon monoxide layer; the first type having thicker layers. The light output from the first type approached 70 nits (20 ft lamberts) at 20 v; and the second type gave the same output at 19–30 v with Al negative and at 60–70 v, with Al positive. Power efficiencies of the order of $4 \times 10^{-3}$% were reported. An interesting feature is that the life was observed to be longer with the Al negative; the time to half initial brightness was of the order of three hours compared with ten minutes with Al positive. Without more details of the thicknesses of the component layers, discussion of the mechanism in terms of copper sulphide is not worth while.

DCEL in ZnS : Mn films containing no copper was reported by Antcliffe (1966) and was attributed to excess zinc. The maximum brightness, however, was only $10^{-11}$ lamberts (or $10^{-8}$ ft lamberts). From the current–voltage characteristics, a voltage-dependent contribution to the film conductivity was found when zinc was present owing to field ionization of zinc donor levels. Investigation of the spectral emission led to the observation of two peaks; the characteristic manganese emission at 5860 Å and a second peak at 5650 Å. This latter peak was only obtained with zinc and its intensity decreased with increasing manganese concentration. A mechanism of energy transfer from zinc to manganese was proposed to account for these observations.

The application of dcel ZnS : Mn, Cu, Cl thin films for visual displays has been reported by King (1967). On the initial application of a d.c. voltage, a marked forming process was observed. The initial high current falls as the brightness builds up in the first few seconds of operation. The light output then decays fairly rapidly reaching half its initial value within about thirty seconds. However, with prolonged operation, the slope of the brightness–time curve decreases with time. By employing stepwise voltage increases, King demonstrated that the brightness could be maintained between 1 and 6 ft lamberts for over a thousand hours. The maximum voltage that could be applied was limited by dielectric breakdown. A maximum brightness of 300 ft lamberts at 50 v was reported, the film having been aged for 10 hours at 20 v. From the observed current of 10 mA and the cell area (1·95 $mm^2$), a power efficiency at the 300 ft lambert level, of the order of $3 \times 10^{-3}$% is indicated.

Soeya (1967) also observed a forming process in ZnS : Mn, Cu, Cl thin films and her results appear to support the oxidation mechanism put forward by Goldberg and Nickerson (1963). With the aluminium electrode positive, forming occurred in air but not in a vacuum ($5 \times 10^{-5}$ torr). Introduction of air into the evacuated vessel gave rise to a fall in current accompanied by the build up of dcel emission. Subsequent evacuation of the vessel did not produce a change in the current. Using gold electrodes, the behaviour of the films was somewhat different: dcel emission was observed at higher voltages and the build-up was much slower. Microscopic observations indicated that the emission for the two types originated from different regions. The films contained a number of microcrystalline areas within the phosphor. With Au electrodes emission appeared from the interfaces between

the crystalline grains and the phosphor. Using Al electrodes, on the other hand, the light output originated from within the phosphor. The blocking (or barrier) layer which determined the site of emission was taken to be aluminium oxide in the case of Al electrodes. With Au, however, no such oxide layer would exist so the blocking layers at the interface within the film were dominant. The nature of the crystalline grains was unspecified by Soeya, but regions of excess copper could account for local recrystallization and give rise to $Cu_xS$/ZnS interfaces suitable for dcel.

A forming process has also been observed on ZnS : Mn, Cu, Cl films by Uchida (1968). Microscopic studies were carried out on sandwich and gap cells, and the results were similar to those reported by Goldberg and Nickerson (1963). Although forming occurred using both aluminium and gold electrodes, the effect was more marked with Al. He concluded that the decrease in current may not only result from the formation of a blocking layer between the ZnS film and the metal electrode (i.e. aluminium oxide), but the migration of excess copper near the film surface may be a contributory factor. The emission appeared to be located near the metal anode and hole injection from a p-type copper compound ($Cu_2O$ or $Cu_2S$) was proposed. In the sandwich cell, emission was observed at less than 30 v only if the Al electrode was positive. At higher voltages dcel was also observed with the Al negative but was much less uniform.

The role of copper sulphide in the dcel of ZnS films was demonstrated by Vlasenko and Gergell (1968). It had been shown earlier that d.c. emission from ZnS films containing only manganese was very low owing to polarization effects (Vlasenko and Zynio 1967). By depositing p-type $Cu_xS$ onto conducting glass followed by a ZnS : Mn, Cu, Cl film, a p–n heterojunction was formed. The forward characteristics of the junction showed typical exponential dependence of current on voltage below about 1 v. A linear relationship was obtained at higher voltages, and the temperature dependence of the slope indicated a Fermi level depth in the ZnS film of 0·2–0·25 ev. The Fermi level in the $Cu_xS$ layer was located in the valence band at between 1·25 and 1·7 ev below the bottom of the conduction band for resistivities between $10^{-1}$ and $10^{-3}$ Ω cm. Rectification was observed and the reverse current was about three orders lower than in the forward direction at 1 v. Dcel was observed with both forward and reverse bias. Higher light outputs were obtained with reverse bias (>30 ft lamberts at 30 v), and the power efficiencies were of the order of $5 \times 10^{-3}$ w/w and $10^{-5}$ w/w in the reverse and forward directions respectively. A mechanism of impact excitation or ionization was proposed to account for the reverse bias characteristics. Electrons penetrate through the thin barrier between the tin oxide conducting glass and the $Cu_xS$, and tunnel from the $Cu_xS$ valence band into the ZnS. Acceleration of these electrons in the depletion region of the ZnS can produce carrier multiplication by ionization of deep donor centres or the lattice, and the manganese centres will be directly excited by impact. The forward bias electroluminescence is assumed to be due to hole injection from $Cu_xS$ into ZnS.

### 3.3. *Single crystals*

Zinc sulphide in single-crystal form offers the best opportunity for studies aimed at elucidating the mechanisms of electroluminescence. Unfortunately a number of contributory factors have resulted in single crystals being the least studied form. Apart from the difficulty in preparing crystals, ZnS cannot be made strongly p-type because of the vacancy compensation so that the production of p–n junctions for efficient injection electroluminescence is a remote possiblity. No quantitative dcel results have been reported on single crystals to date, but some qualitative observations will be briefly reviewed.

Piper and Williams (1952) first achieved light emission by the application of a d.c. potential to a ZnS : Cu single crystal. The light output increased exponentially with voltage and linearly with current. Emission was located at the cathode and a mechanism of electron tunnelling through a potential barrier followed by impact excitation of activator centres was proposed. Under a.c. conditions, low applied voltages gave rise to two light pulses per cycle out of phase with the voltage. Above a critica voltage, additional peaks appeared

which were in phase with the voltage and were attributed to the onset of tunnelling. At lower voltages, the electrons were supplied by field ionization of deep donor levels.

Frankl (1955) also observed 'in phase' and 'out phase' peaks under 60-cycle half-wave excitation. The 'in phase' peak became dominant at high applied voltages (>4 kv) and a power efficiency of the order of $10^{-5}$ w/w was indicated. The emission originated from localized areas which were distributed uniformly through the crystal. Internal barriers due to imperfections caused the field to be concentrated in these localized areas sufficiently to accelerate electrons to impact ionization energies. Later, Frankl (1958) observed emission located at the cathode. This was attributed to more perfect crystals, the internal barriers were now negligible compared with the cathode contact barrier. He found that manganese activated crystals exhibited only 'in phase' electroluminescence.

The observation of d.c. edge emission from ZnS single crystals by Smith (1955) was ascribed to the injection of both holes and electrons. The broad-band emission extended from the absorption edge (3400 Å) to 6000 Å. The crystals were not intentionally doped, but no analysis of impurity content was reported. Neumark (1956) investigated the dcel and thermoluminescence of ZnS single crystals as a function of temperature. The observed increase in d.c. light output with temperature was explained by impact ionization caused by electrons thermally released from traps. Steinberger *et al.* (1961) observed dcel in single crystals of ZnS : Cu, but the emission was some two orders of magnitude below that under a.c. excitation. They attempted to verify the existence of impact ionization using rectangular voltage pulses, but could not distinguish between ionization and electron tunnelling at the cathode barrier. The current and light output varied as $\exp\{-\text{constant}/(\text{volts})^{1/2}\}$ for both d.c. and a.c. excitation.

The first claim of p–n junction electroluminescence was made by Narita (1960). He observed emission from single crystals of ZnS : Cl, which were believed to be n-type, on which copper was evaporated and diffused in to form a p-type region. Under pulse excitation, a brightness peak occurred at the end of the voltage pulse in the forward direction, and at the beginning of the pulse with reverse bias. Aven and Cusano (1964) prepared p–n heterojunctions by epitaxially depositing p-type copper chalcogenides onto ZnS and ZnSe crystals and films. Deposition was carried out by immersion in a hot solution of $Cu^+$ ions. Injection electroluminescence due to hole injection from the chalcogenide was recorded. Some difficulty was experienced in making ohmic contacts to the ZnS crystals, so the major effort was concentrated on $Cu_2Se$/ZnSe junctions. A light emission threshold of 1·4 v was reported at room temperature, and a brightness of 700 ft lamberts (at an estimated quantum efficiency of $2\times10^{-5}$ photons/electron) was achieved at 77 °K. ZnS/$Cu_2S$ heterojunctions were fabricated by Georgobiani and Steblin (1967 a); the copper sulphide, in this case, was evaporated onto the polished and etched ZnS crystal surface. Only reverse bias electroluminescence was observed and the brightness varied as $\exp(-V^{-1/2})$ over four orders of magnitude. However, after heat treatment, forward bias emission was obtained. This was attributed to thermal copper diffusion to produce a p-type ZnS region: the electroluminescence was then considered to take place at a p–n junction in the ZnS crystal. By diffusing copper into n-type ZnS : Cl crystals, Georgobiani and Steblin (1967 b) observed forward bias emission at 2·2 v and reverse bias emission at 18 v. Similar behaviour was reported for silver diffusion into ZnS : Cl crystals; forward and reverse electroluminescence was achieved at 25 v (Georgobiani and Steblin 1967 c).

Low voltage dcel of ZnS : Cu crystals was reported by Indradev and Garlick (1967). The crystals were doped with copper by embedding and firing in a ZnS : Cu powder, and contained $10^{-3}$ g/g copper. At low d.c. voltages, rapidly decaying pulses of light were observed, but above a certain threshold voltage, steady emission was obtained. This threshold was sometimes as high as 60 v, but for one crystal, steady emission was achieved at 6 v, applied perpendicular to the *c* axis. It was suggested that the copper formed precipitates at stacking faults, which behaved as reservoirs of holes. The decay of brightness at low voltages was due to depletion of the holes. At higher voltages, the field was high enough for impact ionization of the copper centres to take place giving rise to steady

emission. Biter *et al.* (1969) achieved dcel at very low voltages in undoped and copper-doped ZnS crystals which had been heat treated in molten zinc. The crystals were heated in zinc at 900 °c for 10–40 hours and quenched in air or water. The undoped crystals exhibited the self-activated blue emission down to 1·6 v d.c. at an estimated quantum efficiency of $10^{-11}$ photons/electron. Green emission was observed from the copper-doped crystals at a threshold voltage of 2 v at a comparable efficiency. The emission occurred at the cathode and was diffuse at low voltages.

As the voltage was increased, the area of the emitting surface decreased, and the emission appeared as bright spots at the cathode above about 15 v. To account for the emission at less than band-gap voltages, the authors proposed the existence of n*–n junctions. The n* regions are strongly conductive and the compensated, less conductive n regions result from the quenching operation. The energy of the emitted photon originates partly from a hot electron ionizing a luminescent centre by an impact-tunnelling transition at an n*–n junction, and partly from a second hot electron which transverses the n* region to be radiatively captured by the ionized luminescent centre in the n-region. The high-voltage characteristics were believed to involve impact ionization in a Schottky barrier at the cathode.

In the last few years, there has been a considerable increase in research into rare-earth doped II–VI compounds. Razi and Anderson (1967) achieved green electroluminescence from terbium-doped ZnS crystals under both a.c. and d.c. pulse excitation. The integrated brightness under pulses followed the $\exp\{-\text{constant}/(\text{volts})^{1/2}\}$ law, giving a four orders of magnitude increase in light output for a rise in peak pulse voltage amplitude from 650 to 1500 v. The estimated efficiencies were approximately $5\times10^{-2}\%$ (power) and 10% (quantum).

## 4. Mechanisms

As Morehead (1967) has summarized, electroluminescence in zinc-sulphide-type phosphors can be produced in at least two ways: by the acceleration of majority carriers so that they gain enough energy to impact ionize the lattice or a filled recombination centre, and subsequently recombine radiatively with minority carriers; or by the injection across a potential barrier of minority carriers which recombine with majority carriers. The recombination, in both cases, may occur directly (i.e., across the band gap) or via recombination centres; the latter being the most unlikely. A third possible mechanism is that of bipolar tunnel injection from inclusions of low band-gap material. Fischer (1963) suggested that copper sulphide inclusions in zinc sulphide behave in this manner.

### 4.1. *Carrier acceleration and impact ionization*

It is thought that high electric fields in excess of $10^5$ v cm$^{-1}$ are required for this mechanism (e.g. Zalm 1956, Piper and Williams 1958). The field will be concentrated in a barrier region produced by carrier exhaustion at, for example, a metal–phosphor contact or a reverse biased p–n junction. The carriers for acceleration can be supplied by tunnelling through the barrier, Schottky emission over the barrier, or by thermal generation in the phosphor material. Since the ionization occurs in a high-field region, the free carriers (predominantly electrons in zinc sulphide) will be swept away from the ionized recombination centres and, under direct current excitation, the recombination rate will be low. Under alternating current excitation, however, some electrons apart from those that may become trapped, will return to the region ionized in the previous half cycle and recombine under low-field conditions. Thus, dcel in ZnS will be inefficient if the mechanism of impact ionization prevails. Figure 1 illustrates this mechanism for the frequently discussed p$Cu_xS$/nZnS heterojunction under reverse bias.

### 4.2. *Minority injection*

The least complicated mechanism is that of minority carrier injection by means of a metal contact or a p–n junction. Thornton (1961 b) proposed this mechanism for dcel in ZnS thin films. He maintained there were stacked p–n junctions connected by conducting

inclusions of copper sulphide, and that, under forward bias, radiative recombination occurs. The main drawback to the application of this mechanism to ZnS is that thermal carrier regeneration is negligible over a band gap of 3·7 ev. Furthermore, owing to vacancy compensation effects, p-type conductivity in ZnS is rather low so that p–n junctions will become so broad that dotted electroluminescent lines should be visible under the microscope. This is not the case, even with the highest resolution (Fischer 1963).

4.3. *Bipolar field emission*

Fischer's model of bipolar field emission (or tunnel injection) again relies on the presence of conducting precipitates of copper sulphide along imperfection lines in insulating ZnS powder particles. The copper sulphide phase acts as a supply of both holes and electrons which will be thermally regenerated. At least one type of carrier will be tunnel injected. Since the conducting inclusions must supply equal amounts of positive and negative charge to prevent charging up, the Fermi level in the copper sulphide corresponds to the middle of the ZnS band gap and the tunnelling threshold will be less than half the band gap. The

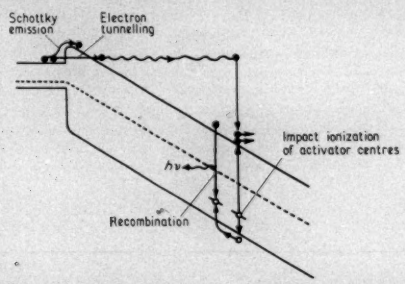



Figure 1. Schematic representation of p$Cu_xS$–nZnS heterojunction under reverse bias.

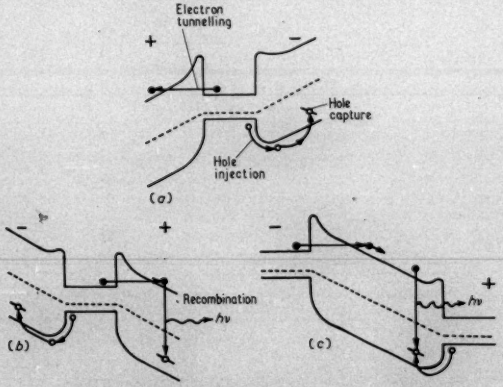



Figure 2. Bipolar field emission from $Cu_xS$ into ZnS.

injected holes are rapidly captured at activator centres and light is emitted when free electrons drop into these centres. Although this model was originally proposed to explain the electroluminescence observed under alternating current conditions, it can also be applied to direct current excitation. In Fischer's model, holes are captured at centres near the cathode end of a conducting inclusion (figure 2(*a*)), and, on field reversal, light is emitted when electrons recombine with the trapped holes (figure (2*b*)). These electrons are either field-emitted at the same end of the inclusion (now the anode), or are field released from traps at the other end of the inclusion. For d.c. operation, two adjacent inclusions are required, so that one supplies holes and the other supplies electrons. Recombination will take place in the ZnS region separating the inclusions, and will be located close to the one that is nearest the external anode because of the fast trapping of holes in ZnS (figure 2(*c*)). Trapping of electrons before they recombine with the trapped holes will be a source of inefficiency.

### 4.4. *The role of manganese*

We will now consider the unique properties of manganese (or rare-earth)-type luminescent centres in ZnS. Manganese usually exists as $Mn^{2+}$ and at least five excited states have been observed from optical absorption measurements (McClure 1963). The position of the ground state has not yet been satisfactorily confirmed, but is most commonly predicted to be well below the valence band edge (Gumlich *et al.* 1966). Excitation of the manganese centre can be achieved by resonance transfer or by direct excitation. Briefly, the resonance transfer mechanism involves radiationless transfer of energy from a sensitizer centre (in our case, copper) to the main activator centre (manganese). This energy raises electrons from the ground state to an excited state and emission occurs when the electrons drop back

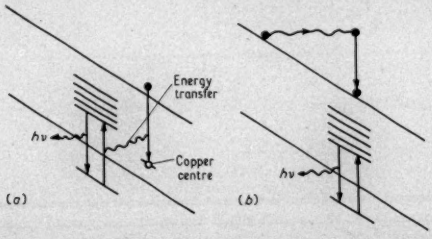



Figure 3. Excitation of the manganese centre: (*a*) resonance energy transfer, (*b*) direct excitation.

again (figure 3(*a*)). Shionoya (1956) maintained that this process was dominant in ZnS (Mn, Cu) powder from optical measurements. The physical details of the transfer are rather complicated, and are believed to involve overlap of the electric dipole fields of the sensitizer and activator (Dexter 1953). Direct excitation of the manganese centres arises by acceleration of electrons which collide with the centres and excite them without ionization (figure 3(*b*)). The significant difference between this and impact ionization is that the recombination within the isolated manganese centre will not be affected by external field conditions. The recombination probability may be expected to be high even in high-field regions.

### 4.5. *Luminance–current–voltage characteristics*

Any proposed mechanism of dcel must account for the observed luminance–current–voltage characteristics. Most studies have, of course, been carried out with a.c. excitation,

but, as illustrated by Zalm (1956), dcel emission from ZnS–$Li_2S$–Mn followed the same voltage dependence as a.c. phosphors.

The high-field type of electroluminescence can be attributed to field emission (or tunnelling) in barrier regions, or impact ionization. Unfortunately, both processes can give rise to the same characteristics. If the barrier is formed at the cathode by the ionization of donors under the action of the applied field, the exhaustion layer formed concentrates the field at the cathode. Tunnelling of electrons from the contact gives a current proportional to $\exp(-\text{constant}/V^{1/2})$. (This follows from the Fowler–Nordheim field-emission equation, the field being proportional to $V^{1/2}$.) The probability that these electrons will be excited to sufficient energies to ionize the lattice or the luminescent centres, follows the same exponential dependence on voltage, but the constants of proportionality will be different. Alternatively, the cathode barrier may be due to a depletion of donor centres at the phosphor surface. The field is then proportional to $V$ and the tunnel current is proportional to $\exp(-\text{constant}/V)$. The pre-exponential terms are also dependent on voltage; being proportional to $V$ and $V^2$ for the exhaustion and depletion layers respectively. The exponent will, however, be dominant. The light emission is determined by the number of ionized luminescent centres so the luminance–voltage characteristic will obey a similar law.

This theory, developed notably by Piper and Williams (1955), Alfrey and Taylor (1957) and Zalm (1956), is limited to a single barrier at the cathode. As Frankl (1958) pointed out, evidence for internal barriers in ZnS single crystals and powders does exist. These give rise to emission from discrete spots, observed in crystals by Diemer (1955) and in powders by Waymouth and Bitter (1954). Lehmann (1960) maintained that the a.c. luminance–voltage characteristic of a single phosphor particle had the form, $L=L_0\exp(-V_0/V)$, but because of a broad particle-size distribution, the characteristic of a powder layer followed the familiar $V^{-1/2}$ law. He supported this by visual observations on single particles and by mathematical analysis. The presence of conducting inclusions of copper sulphide is important in this context. If a number of barriers exists within a single particle due to stacking faults, etc., the decoration of these with copper sulphide will assist the tunnelling of electrons by reduction of the barrier height. One would expect that, as more copper sulphide is added, the number of active barriers will increase to give a more uniform light source.

As we have already discussed, minority carrier injection at forward-biased p–n junctions is unlikely in ZnS. The current–voltage characteristics would have the form $I=I'\exp\{(V/V')-1\}$, and the luminance should be proportional to the square of the current since the recombination is a two-stage, bimolecular process. Nevertheless, Thornton (1962) reported these characteristics in powders and thin films under d.c. excitation.

The incorporation of both manganese and copper adds to the complexity and makes interpretation even more difficult. If the resonance transfer mechanism is operative, excitation will occur at the copper centre and the current–voltage characteristics will be determined by tunnelling, impact ionization, or both. The transfer of energy and subsequent recombination at the manganese centre will be a bimolecular process (Curie 1960), so that the luminance should follow a square-law dependence on current. The luminescence will then have the same exponential dependence on voltage as the current, but the exponent will be doubled.

Direct excitation of the manganese centre by accelerated electrons does not involve impact ionization, but it is still likely to occur at other lattice sites in the high fields that are present. For this reason, or because the electrons are supplied by tunnelling, the current will still be proportional to $\exp(-\text{constant})/V^{1/2}$. The recombination in this case is monomolecular so that a linear luminance–current characteristic should be observed. Mattler and Ceva (1962) observed two separate straight-line regions in their logarithm (luminance) against $(\text{voltage})^{-1/2}$ plots on ZnS (Mn, Cu) under a.c. excitation. They attributed this to a changeover from energy transfer (at the low-voltage end) to direct excitation of the manganese centres; the reason for the increase in slope being the greater ionization energy of the manganese compared with that of the copper centre. However, the position of the manganese centre in the ZnS band picture is still a matter for conjecture, but it is believed to be too far below the conduction band to be ionized.

### 4.6. *Efficiency*

Most estimates of the theoretical maximum efficiency of electroluminescence have been made for the case of a.c. excitation. The main light peak then occurs on field reversal as electrons travel back to the ionized region. Morehead (1960) arrived at a figure of 38 lm $w^{-1}$ for an insulating powder phosphor, in which the dominant mechanism is impact ionization. The highest experimental value reported is 14 lm $w^{-1}$ for a green-emitting ZnS : Cu, Cl powder cell (corresponding to 18–19 lm $w^{-1}$ for the phosphor itself) under a.c. excitation (Lehmann 1958). In other words, nearly 5% of the dissipated energy is converted into light. With d.c. bias, the recombination must occur in a high-field region, except in the case of minority carrier injection. This will reduce the probability of recombination; an estimate of between $10^{-7}$ and $10^{-9}$ was made by Haake (1957) in a field of $10^6$ v $cm^{-1}$. Thus, dcel involving impact ionization will have a low efficiency.

Forward-injection electroluminescence is a much better prospect for high efficiency under d.c. conditions. Fischer (1961) maintains that a p–i–n arrangement offers the most promise. Owing to the absence of suitable p-type materials with the required high band gap, the incorporation of an insulating region effectively isolates the semiconducting and luminescent parts of the device. The i-region can be doped for optimum recombination efficiency, and injection takes place via diffusion or space-charge currents. An overall efficiency of 10% w/w is estimated and "there is much space left for improvements". Furthermore, if 40 w $cm^{-2}$ of Joule heat can be tolerated, a brightness of nearly $1{\cdot}5 \times 10^6$ ft lamberts may be obtained for green light.

Manganese-activated phosphors do not suffer a loss in efficiency if the recombination takes place in a high field. The $Mn^{2+}$ centres are isolated and every electron that acquires sufficient energy from the applied field to collision-excite the centre will produce a photon. It is thought (Morehead 1960) that this direct excitation mechanism only dominates at high currents. The efficiency of the process will be appreciably higher than that of impact ionization. Cusano (1967) maintains that for most manganese activated electroluminescent structures, the excitation occurs at reverse-biased copper sulphide–zinc sulphide p–n junctions. Majority carrier (electron) injection (by tunnelling) into a high-field region in the zinc sulphide occurs, and a fraction of these electrons are accelerated to energies sufficient to impact-excite the manganese ion. Efficiencies of $2 \times 10^{-4}$ w/w for thin films, in which this mechanism was believed to dominate, have been reported (Cusano 1962).

We also consider direct collision excitation of manganese centres to be the dominant mechanism in our powder layers, and the importance of a copper sulphide phase has been demonstrated (Vecht *et al.* 1969). The high efficiencies achieved ($\sim 3{\cdot}3 \times 10^{-3}$ w/w) are consistent with direct excitation and recombination within the isolated manganese centres.

## 5. Future developments

Up to the present time, only a small proportion of the study of electroluminescence has been devoted to dcel of manganese-activated II–VI compounds. It is hoped that this review indicates that there is considerable promise in this area and warrants further extensive investigation.

The maintenance characteristics under continuous operation must be improved considerably. A life of several thousand hours is required before useful display devices can be made. Apart from the effects of moisture, control of the stoichiometry of the electroluminescent material may well give rise to more stable performance. Lehmann (1966) has demonstrated that firing a ZnS powder phosphor in a sulphur atmosphere gave improved life under a.c. excitation. It is thought that this treatment prevented the migration of sulphur vacancies. The migration of other atomic or ionic species should also be investigated. The forming process, observed in powders and thin films, may well be of great significance in this context.

In addition to the optimization of the performance of manganese-doped materials, the incorporation of rare earths offers exciting possibilities. Provided rare-earth ions can be

introduced into the ZnS lattice and behave in a similar manner to manganese, a range of different colours becomes available.

**References**


ADDAMIANO, A., and AVEN, M., 1960, *J. Appl. Phys.*, **31**, 36–9.

ALFREY, G. F., and TAYLOR, J. B., 1955, *Brit. J. Appl. Phys.* (Suppl.), **4**, S44–9.

ALFREY, G. F., and TAYLOR, K. N. R., 1957, *Helv. Phys. Acta*, **30**, 206–8.

ANTCLIFFE, G. A., 1966, *Brit. J. Appl. Phys.*, **17**, 327–36.

AVEN, M., and CUSANO, D. A., 1964, *J. Appl. Phys.*, **35**, 606–11.

AVEN, M., and PARODI, J. A., 1960, *J. Phys. Chem. Solids*, **13**, 56–64.

BITER, W. J., INDRADEV and WILLIAMS, F., 1969, *J. Phys. Chem. Solids*, **30**, 503–8.

BLANK, Y. S., and YATLINKO, I. I., 1964, *Sov. Phys.–Solid St.*, **5**, 1963–5.

BOWTELL, J. N., and BATE, H. C., 1956, *Proc. Instn Radio Engrs*, **44**, 697.

BUTLER, K. H., and WAYMOUTH, J. F., 1955, *Brit. J. Appl. Phys.* (Suppl), **4**, S33–7.

CURIE, D., 1960, *Luminescence in Crystals* (Paris: Dunod).

CUSANO, D. A., 1962, *Luminescence of Organic and Inorganic Materials*, Eds. H. P. Kallman and G. M. Spruch (New York: John Wiley), pp. 494–522.

——1967, *Physics and Chemistry of II–VI Compounds*, Eds. M. Aven and J. S. Prener (Amsterdam: North-Holland), pp. 709–66.

DESTRIAU, G., 1936, *J. Chim. Phys.*, **33**, 587.

DEXTER, D. L., 1953, *J. Chem. Phys.*, **21**, 836–50.

DIEMER, G., 1955, *Philips Res. Rep.*, **10**, 194–204.

DRESNER, J., and SHALLCROSS, F. V., 1963, *J. Appl. Phys.*, **34**, 2390–5.

FAVORIN, V. N., and KOZINA, G. S., 1961, *Opt. Spectrosc.*, **10**, 43–5.

FISCHER, A. G., 1961, *Solid-St. Electron.*, **2**, 232–46.

—— 1963, *J. Electrochem. Soc.*, **110**, 733–48.

—— 1966, *Luminescence of Inorganic Solids*, Ed. P. Goldberg (New York: Academic Press), pp. 541–602.

FISCHER, A. G., SHRADER, R. E., and LARACH, S., 1965, *Photoelectronic Materials and Devices*, Ed. S. Larach (New York: Van Nostrand) pp. 1–99.

FRANKL, D. R., 1955, *Phys. Rev.*, **100**, 1105–11.

—— 1958, *Phys. Rev.*, **111**, 1540–9.

FRIDKIN, V. M., 1961, *Opt. Spectrosc.*, **11**, 54–8.

GEORGOBIANI, A. N., and STEBLIN, V. I., 1967 a, *Sov. Phys.–Semiconductors*, **1**, 772–4.

—— 1967 b, *Phys. Stat. Sol.*, **21**, K49–51.

—— 1967 c, *Fiz. Tekh. Poluprov.*, **1**, 956–7.

GILLES, J. M., and VAN CAKENBERGHE, J., 1960, *Solid State Physics in Electronics and Telecommunications*, Eds. M. Désirant and J. L. Michiels, Vol. 2 (London: Academic Press), pp. 900–6.

GOLDBERG, P., and NICKERSON, J. W. 1963, *J. Appl. Phys.*, **34**, 1601–8.

GOLOVKINA, E. D., PASYNKOV, V. V., and KHANINA, G. V., 1965, *Opt. Spectrosc.*, **19**, 158–9.

GREENE, L. C., REYNOLDS, D. C., CZYZAK, S. J., and BAKER, W. M., 1958, *J. Chem. Phys.*, **29**, 1375–80.

GUMLICH, H. E., PFROGNER, R. L., SHAFFER, J. C., and WILLIAMS, F. E., 1966, *J. Chem. Phys.*, **44**, 3929–34.

HAAKE, C. H., 1957, *J. Electrochem. Soc.*, **104**, 291–8.

HALSTEAD, R. E., and KOLLER, L. R., 1954, *Phys. Rev.*, **93**, 349–50.

HENISCH, H. K., 1962, *Electroluminescence* (London: Pergamon).

—— 1964, *Rep. Prog. Phys.*, **27**, 369–405.

INARI, T., and TAKEDA, T., 1964, *Jap. J. Appl. Phys.*, **3**, 802–3.

INDRADEV and GARLICK, G. F. J., 1967, *Brit. J. Appl. Phys.*, **18**, 1–6.

IVEY, H. F., 1963, *Advances in Electronics and Electron Physics* (Suppl. 1)—*Electroluminescence and Related Effects* (London: Spon).

JACKSON, P. A., and VECHT, A., 1970, *R.I.C. Rev.*, to be published.

KAZANKIN, O. N., PEKERMAN, F. M., and PETOSHINA, L. M., 1959, *Opt. Spectrosc.*, **7**, 458–9.

KING, R. C., 1967, *Dallas Meeting, Electrochemical Society, May* 1967.

KREMHELLER, A., 1955, *Sylvania Tech.*, **8**, 11–5.

LEHMANN, W., 1958, *J. Electrochem. Soc.*, **105**, 585–8.

—— 1960, *J. Electrochem. Soc.*, **107**, 20–6.

—— 1966, *J. Electrochem. Soc.*, **113**, 40–2.

LEHOVEC, K., ACCARDO, C. A., and JAMGOCHIAN, E., 1951, *Phys. Rev.*, **83**, 603–7.
LEVERENZ, H. W., 1950, *An Introduction to Luminescence of Solids* (New York: John Wiley), pp. 473–6.
LOSSEV, O. W., 1923, *Telegrafia i Telefonia*, **18**, 61.
MATTLER, J., and CEVA, T., 1962, *Luminescence of Organic and Inorganic Materials*, Eds. H. P. Kallman and G. M. Spruch (New York: John Wiley), pp. 537–48.
MCCLURE, D. S., 1963, *J. Chem. Phys.*, **39**, 2850–5.
MOREHEAD, F. F., 1960, *J. Electrochem. Soc.*, **107**, 281–7.
—— 1967, *Physics and Chemistry of II–VI Compounds*, Eds. M. Aven and J. S. Prener (Amsterdam: North-Holland), pp. 613–56.
NARITA, S., 1960, *J. Phys. Soc. Japan*, **15**, 128–36.
NEUMARK, G. F., 1956, *Phys. Rev.*, **103**, 41–6.
NICKERSON, J. W., and GOLDBERG, P., 1963, *Trans. Natn. Vac. Symp.*, **10**, 475–9.
NICOLL, F. N., and KAZAN, B., 1955, *Proc. Instn Radio Engrs*, **43**, 1012.
PATEK, K., 1961, *Czech. J. Phys.*, **B11**, 18–20.
PIPER, W. W., 1952, *J. Chem. Phys.*, **20**, 1343.
PIPER, W. W., and WILLIAMS, F. E., 1952, *Phys. Rev.*, **87**, 151–2.
—— 1955, *Brit. J. Appl. Phys.* (Suppl.), **4**, S39–44.
—— 1958, *Solid State Physics, Advances in Research and Applications*, Eds. F. Seitz and D. Turnbull, Vol. 6 (New York: Academic Press), pp. 95–173.
RAZI, S., and ANDERSON, W. W., 1967, *Trans. Metall. Soc. AIME*, **239**, 350–4.
SHIONOYA, S., 1956, *Bull. Chem. Soc., Japan*, **29**, 935–44.
SMITH, R. W., 1955, *Phys. Rev.*, **100**, 760–1.
SOEYA, T., 1967, *Jap. J. Appl. Phys.*, **6**, 205–13.
STEINBERGER, I. T., BAR, V., and ALEXANDER, E., 1961, *Phys. Rev.*, **121**, 118–24.
STUDER, F. J., and CUSANO, D. A., 1955, *J. Opt. Soc. Amer.*, **45**, 493–7.
TE VELDE, T. S., 1964, *Electrochem. Abstr.*, **13**, 70–3.
THORNTON, P. R., 1967, *The Physics of Electroluminescent Devices* (London: Spon).
THORNTON, W. A., 1958, *Bull. Amer. Phys. Soc.*, **3**, 233.
—— 1959 a, *Phys. Rev.*, **113**, 1187–91.
—— 1959 b, *J. Appl. Phys.*, **30**, 123–4.
—— 1959 c, *Phys. Rev.*, **116**, 893–4.
—— 1961 a, *Phys. Rev.*, **122**, 58–9.
—— 1961 b, *J. Electrochem. Soc.*, **108**, 636–45.
—— 1962, *J. Appl. Phys.*, **33**, 3045–8.
—— 1963, *Electrochem. Abstr.*, **12**, 59–63.
UCHIDA, W., 1968, *Jap. J. Appl. Phys.*, **7**, 378–85.
VECHT, A., 1966, *Physics of Thin Films*, Eds. G. Hass and R. E. Thun, Vol. 3 (London: Academic Press), pp. 165–210.
—— 1969, *J. Vac. Sci. Technol.*, **6**, 773–6.
VECHT, A., and ELLIS, R., 1966, *Nature, Lond.*, **210**, 1251–2.
—— 1968, *Proc. Int. Conf. on Luminescence*, 1966 (Budapest: Publishing House of the Hungarian Academy of Sciences), pp. 1986–91.
VECHT, A., and ELY, B. W., 1964, *Electrochem. Abstr.*, **13**, 76–8.
VECHT, A., WERRING, N. J., ELLIS, R., and SMITH, P. J. F., 1969, *J. Phys.* D: *Appl. Phys.* **2**, 953–66.
VECHT, A., WERRING, N. J., and SMITH, P. J. F., 1968, *J. Phys.* D: *Appl. Phys.* **1**, 134–6.
VERESHCHAGIN, I. K., 1962, *Opt. Spectrosc.*, **12**, 424–9.
VLASENKO, N. A., and GERGELL, A. N., 1968, *Phys. Stat. Sol.*, **26**, K77–81.
VLASENKO, N. A., and ZYNIO, S. A., 1967, *Phys. Stat. Sol.*, **20**, 311–20.
WAYMOUTH, J. F., and BITTER, F., 1954, *Phys. Rev.*, **95**, 941–9.
ZALLEN, R., ERIKSEN, W. T., and AHLBURG, H., 1960, *J. Electrochem. Soc.*, **107**, 288–95.
ZALM, P., 1956, *Philips Res. Rep.*, **11**, 353–99, 417–51.
ZALM, P., DIEMER, G., and KLASENS, H. A., 1954, *Philips Res. Rep.*, **9**, 81–108.
—— 1955, *Philips Res. Rep.*, **10**, 205–15.

E. H. RHODERICK

Department of Electrical Engineering and Electronics,
University of Manchester Institute of Science and Technology

# The physics of Schottky barriers†

**Abstract.** A review is given of the physical processes which determine the height of the barrier and the current–voltage relationship in a metal–semiconductor Schottky barrier.

## 1. Historical

The metal–semiconductor rectifier must be the oldest solid state device used in electronics, because the first paper published on the subject appeared in 1874, when Braun reported the asymmetrical nature of conduction between metal points and crystals such as lead sulphide. Its application as a radio-frequency detector is almost as old as wireless telegraphy itself, for it is now fairly certain that Branly's 'coherer' must have relied for its action on the rectifying properties of metal contacts separated by an oxide film. In 1906 Pickard took out a patent for a point-contact rectifier using silicon, and in 1907 Pierce published rectification characteristics of diodes made by sputtering metals on to a variety of semiconductors. The point-contact rectifier or 'catswhisker' was used extensively in the early days of radio, but the first real scientific study of the device (and indeed the beginning of semiconductor physics) was stimulated by the wartime use of silicon and germanium point-contact rectifiers as microwave detectors. Point-contact rectifiers were very variable and unreliable in their characteristics, and our present understanding of Schottky barriers has come with the realization that metal films evaporated onto single-crystal semiconductor surfaces under conditions of high cleanliness can show almost ideal rectification characteristics. An extensive account of the early history of metal–semiconductor contacts can be found in Henisch's book (Henisch 1957), and an up-to-date review of the properties and applications of Schottky barriers has been given by Atalla (1967).

## 2. Formation of the barrier

### 2.1. *Simple Schottky theory*

The rectifying properties of a Schottky barrier arise from the presence of an electrostatic barrier between metal and semiconductor. To see how this arises, consider a piece of metal and a piece of semiconductor, both electrically neutral and separated from each other. The energy band diagram is as shown in figure 1(*a*), the zero of energy being the energy of an electron at rest outside either solid. For definiteness we show the case of an n-type semiconductor whose work function is less than that of the metal, which is in practice the most important case. (The argument can be inverted for the case of p-type semiconductors). If the metal and semiconductor are connected electrically by means of a wire, electrons pass from the semiconductor to the metal and the Fermi levels are forced into coincidence. The energy of an electron at rest outside the surfaces of the two solids is no longer the same, and there is an electric field in the gap (as shown in figure 1(*b*) the field is from right to left). There must be a negative charge on the surface of the metal, balanced by a positive charge on the semiconductor. Since the semiconductor is n-type, this positive charge is provided by conduction electrons receding from the surface, leaving uncompen-

† Based on an invited paper given at the third Solid-State Devices Conference, Exeter, 1969.

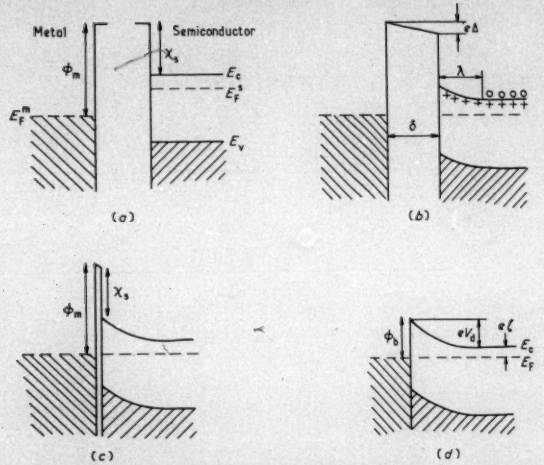



Figure 1. Formation of Schottky barrier from metal and semiconductor. (*a*) neutral and isolated, (*b*) electrically connected, (*c*) separated by narrow gap, (*d*) perfect contact.
○ denotes electron in conduction band; + denotes positive donor ion.

sated donor ions. Because the concentration of these donors is many orders of magnitude less than the concentration of electrons in the metal, the balancing positive charges occupy a layer of appreciable thickness, and the bands in the semiconductor are bent upwards as shown. The work-function of the metal ($\phi_m$) and the electron affinity of the semiconductor ($\chi_s$) remain unaltered as shown. As the metal and semiconductor approach each other, the drop in electrostatic potential ($\Delta$) associated with the field in the gap tends to zero if the field is to remain finite, as shown in figure 1(*c*). Finally, if the gap becomes truly zero (corresponding to perfectly intimate contact between metal and semiconductor), the thin potential barrier separating them disappears altogether, as in figure 1(*d*), and we are left with a barrier arising from the band-bending. Even if contact is not perfect and a very thin high barrier separating metal and semiconductor still remains as in 1(*c*), this barrier is usually so thin ($\sim 10$ Å) that electrons can easily pass through it by tunnelling, and the situation shown in figure 1(*c*) is essentially equivalent to 1(*d*). Clearly, in the limit of perfect contact, $\phi_b = \phi_m - \chi_s$, a relationship which was first derived by Schottky (1938).

### 2.2. *Influence of surface states*

It was soon found that in the case of point contact rectifiers the barrier is often independent of the metal, which is at variance with the simple Schottky theory. Bardeen (1947) was the first to point out the importance of surface states in the semiconductor in determining the barrier height. For our present purpose it is sufficiently accurate to regard surface states as unsaturated (or 'dangling') bonds on the surface of the semiconductor. The surface states are usually continuously distributed in energy within the forbidden gap, and are characterized by a 'neutral level' $\phi_0$ such that if the surface states are occupied up to $\phi_0$ and empty above $\phi_0$ the surface is electrically neutral. If we approximate the Fermi–Dirac distribution by a step function, it follows that when the Fermi level $E_F$ lies above $\phi_0$

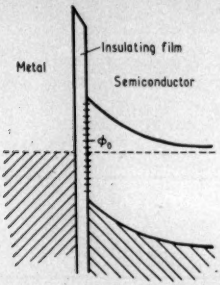



Figure 2. Metal–semiconductor contact with surface states.

the surface states possess a net negative charge, while if $E_F$ lies below $\phi_0$ they possess a net positive charge. In other words, the states below $\phi_0$ are 'donor-like' (positive when empty) and the states above $\phi_0$ are 'acceptor-like' (negative when occupied). In the case shown in figure 2, for which $\phi_0 > E_F$, there is a net positive charge in the surface states, and some of the lines of force terminating on the metal emanate from surface charges rather than from ionized donors; in this case the space charge region is not as thick as when surface states are absent, and the barrier height is reduced. In the same way, if $\phi_0 < E_F$, there is a net negative charge in the surface states and the barrier height is increased. Thus the charge in the surface states has a 'negative feed-back' effect which always tends to reduce the deviation of $\phi_0$ from $E_F$. If the density of surface states is very high, the gain in the feedback loop is very large, and $\phi_0$ tends to be locked to $E_F$ so that the barrier height is virtually independent of the metal. Cowley and Sze (1965) have shown that, in the case of an n-type semiconductor, Bardeen's theory can be approximately expressed in the form

$$\phi_b = \gamma(\phi_m - \chi_s) + (1-\gamma)(E_g - \phi_0) \tag{1}$$

where $E_g$ is the energy gap of the semiconductor and the neutral level $\phi_0$ is measured from the top of the valence band. The approximations leading to (1) are strictly justified only if the charge in the surface states is very much greater than the charge in the space-charge layer. However, in practice it seems to be a reasonable approximation in all cases. This equation assumes that the metal and semiconductor do not make perfect contact as in figure 1(*d*) but remain separated by a thin oxide film of thickness $\delta$ as in figure 1(*c*). $\gamma$ is given by

$$\gamma = \epsilon_i \epsilon_0 / (\epsilon_i \epsilon_0 + e\delta D_s)$$

where $\epsilon_i$ is the relative permittivity of the oxide film, $\epsilon_0$ is the permittivity of free space, and $D_s$ is the density of surface states per electron-volt per unit area of semiconductor surface. If there are no surface states, $\gamma = 1$ and (1) reduces to the simple Schottky relationship $\phi_b = \phi_m - \chi_s$. On the other hand, if $D_s$ is very large, $\gamma$ tends towards zero and $\phi_b$ approximates to $E_g - \phi_0$, i.e. the barrier becomes stabilized so that the Fermi level coincides with $\phi_0$ independently of the value of $\phi_m$.

## 3. Measurements of barrier heights

### 3.1. *Experimental methods*

There are several methods at our disposal for measuring $\phi_b$. The best known relies on the measurement of the differential capacitance $C$ as a function of reverse bias $V$. It can be shown that $C$ is proportional to $(V+V_d)^{-1/2}$, where $V_d$ is the 'diffusion' voltage (i.e. the band bending in volts), so that a plot of $C^{-2}$ against $V$ gives a straight line with

intercept $-V_d$. The barrier height $\phi_b$ is equal to $e(V_d+\zeta)$, where $e\zeta$ is the energy of the conduction band relative to the Fermi level (see figure 1). Alternatively we can measure the current–voltage characteristic and, provided it is close to the ideal rectifier relationship $J=J_0\{\exp(eV/kT)-1\}$, we can write $J_0=AT^2\exp(-\phi_b/kT)$, where $A$ is the Richardson constant corrected for the effective mass of the electrons in the semiconductor (Crowell 1965), so that $\phi_b$ can be obtained from $J_0$. Finally, we can measure the threshold for optical excitation of electrons from the metal over the barrier into the semiconductor. All these methods are intrinsically capable of good accuracy, but the ultimate error is limited by repeatability from sample to sample and is determined by the technology of specimen preparation.

### 3.2. *Results for silicon*

Following the pioneering work of Archer and Atalla (1963), the semiconductor we know most about as far as Schottky barriers are concerned is silicon. An extensive study of metal–silicon barriers by Turner and Rhoderick (1968) has shown that

(i) $\phi_b$ depends on the method of preparation of the surface,

(ii) $\phi_b$ often ages with time, becoming stable in a matter of days or even weeks (see figure 3),

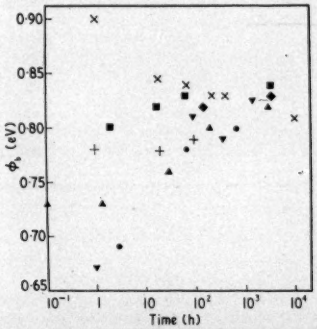



Figure 3. Ageing of gold–silicon Schottky barriers (Turner and Rhoderick 1968). The letters A–E refer to different methods of surface preparation.

| Sample | Resisitivity (Ω cm) | Surface | Vacuum (torr) |
|---|---|---|---|
| × | 1 | A | $10^{-5}$ |
| + | 1 | A | $10^{-7}$ |
| ◆ | 10 | B | $10^{-5}$ |
| ● | 10 | C | $10^{-5}$ |
| ■ | 10 | C | $10^{-5}$ |
| ▲ | 10 | D | $10^{-5}$ |
| ▼ | 10 | E | $10^{-7}$ |

(iii) in the case of etched surfaces the barrier depends on the choice of metal, but not to the extent predicted by the Schottky theory, and

(iv) in the case of surfaces cleaved in ultra-high vacuum, the barrier is almost independent of the metal (see figure 4).

It is clear that cleaved surfaces belong to the case where a high density of states locks the barrier so that it is independent of $\phi_m$ ($\gamma\rightarrow 0$), and that etched surfaces lie somewhere

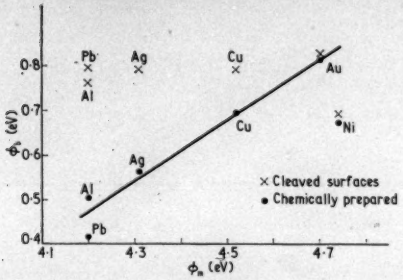



Figure 4. Barrier heights of metal–silicon Schottky barriers (Turner and Rhoderick 1968).

between the two extreme cases with $\gamma \simeq 0{\cdot}67$. The cleaved-surface data are compatible with $D_s \simeq 4 \times 10^{18}\ \mathrm{m^{-2}\,eV^{-1}}$, which corresponds to a total density of surface states of about 1 per atom, as we might expect from a dangling bond picture. The etched-surface data are compatible with $D_s \simeq 10^{16}\ \mathrm{m^{-2}\,eV^{-1}}$, which is a figure often measured on 'real' surfaces by field effect measurements. One can suppose that the thin oxide layer, which always covers etched surfaces, almost completely saturates the dangling bonds.

Any attempt at a very detailed comparison between theory and experiment is hardly profitable because of the uncertainty in the values of the metal work functions. These are generally taken from published data and usually refer to experimental conditions quite different from those under which the metal film has been evaporated. From time to time the accepted values of work functions may be drastically revised; for example, the work function of gold has apparently increased by 0·5 eV in the last few years! Most of the published analyses of Schottky barrier data should be treated with caution. The ideal procedure would be to make a measurement of the semiconductor work function, then evaporate the metal and measure its work function, then measure the height of the Schottky barrier, all in an ultra-high vacuum system without breaking the vacuum. Such a programme has been recently initiated by Poole (1969) but there is not yet sufficient data for any extensive conclusions to be drawn. Preliminary measurements suggest that it may not prove possible to assign to every metal or semiconductor a unique work function which determines in a consistent manner the height of the barrier formed by that material in conjunction with some other semiconductor or metal.

### 3.3. *Results for gallium arsenide*

The next most widely studied semiconductor is gallium arsenide, and here the results are different from silicon in two respects (figure 5). Firstly, on cleaved (110) surfaces there is much more variation with choice of metal than in the case of silicon, suggesting that the surface-state density is lower by about an order of magnitude ($D_s \simeq 7 \times 10^{17}\ \mathrm{m^{-2}\,eV^{-1}}$), and secondly, while there is more dependence on $\phi_m$ in the case of etched surfaces than in the case of cleaved surfaces, the difference is not nearly as marked as in the case of silicon. In fact, the data on etched surfaces are compatible with a surface state density $\simeq 3 \times 10^{17}\ \mathrm{m^{-2}\,eV^{-1}}$, which is only a factor of two lower. The first result is compatible with the theoretical calculations of Jones (1968), and also with the observation by Van Laar and Seheer (1967) that the Fermi level at the free surface of cleaved gallium arsenide surfaces is *not* clamped by surface states as in the case of Si. The second may be due to the fact that gallium arsenide does not oxidize as easily as silicon.

There is also a dependence of barrier height on surface orientation. In the case of gallium arsenide, the $(\bar{1}\bar{1}\bar{1})$ surface gives a value of $\phi_b$ about 0·1 eV higher than the (100)

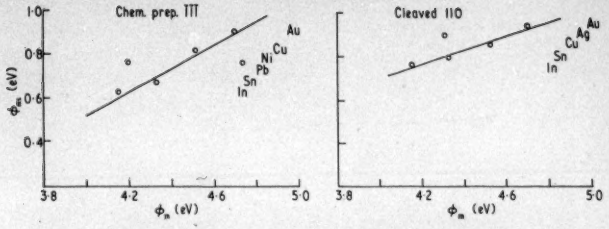



Figure 5. Barrier heights of metal–gallium arsenide Schottky barriers (Smith 1969).

face, and the (111) face is intermediate between the two. The difference between the (111) and ($\bar{1}\bar{1}\bar{1}$) faces is to be expected in polar compounds.

It turns out that, for both silicon and gallium arsenide, $\phi_0 \simeq \frac{1}{3}E_g$. It was suggested by Mead and Spitzer (1964) that this may be true for all semiconductors with the diamond or zincblende structure. Such data as are available suggest that the rule is obeyed quite well—certainly by germanium, silicon, and the III–V compounds.

### 3.4. *Other semiconductors*

Measurements of varying degrees of completeness on a large number of semiconductors have been summarized by Mead (1966) and Sze (1969). Mead suggests that semiconductors can be divided into two categories—those in which the barrier height is controlled by surface states, and those in which surface states are unimportant. It is clear from the foregoing that this is too simple a view, and that some semiconductors (e.g. silicon) may be in either category according to the method of surface preparation. There are some cases, however, in which it is probably legitimate to characterize semiconductors in this way. Mead reports that in the case of cadmium sulphide surface states are unimportant, while for cadmium selenide the barrier height is effectively locked by surface states. In the case of Cd(S, Se) mixed crystals there is apparently a smooth variation from one type of behaviour to the other as the composition is varied from CdS to CdSe.

### 3.5. *p-type semiconductors*

The foregoing results refer exclusively to n-type semiconductors. There is relatively little data on p-type semiconductors, partly because the barriers are so low that measurement is difficult, and partly because there is less technical interest in Schottky diodes made on p-type than on n-type semiconductors as a result of the lower barrier height and lower carrier mobility. The most extensive measurements on p-type semiconductors are those of Smith and Rhoderick (1970) on p-type silicon. Smith has shown that, for barriers on etched p-type silicon, the barrier heights can be explained in terms of the same surface parameters (density of surface-states and position of the neutral level) as in the case of Turner's results on n-type silicon.

It might be expected that the surface properties should not depend on the bulk impurity. For the case of p-type semiconductors, the analogue of equation (1) is

$$\phi_b{}^p = \gamma(\chi_s - \phi_m + E_g) + (1-\gamma)\phi_0. \tag{2}$$

From (1) and (2) we find $\phi_b{}^n + \phi_b{}^p = E_g$, i.e. for a given metal the sum of the barriers on n-type and p-type specimens of the same semiconductor should equal the band-gap, provided the surface-state parameters $\gamma$ and $\phi_0$ are the same in both cases. Table 1 shows that the data for silicon satisfy this relationship quite well.

**Table 1. Experimental barrier height data on etched p-type and n-type silicon. (n-type from Turner and Rhoderick (1968), p-type data from Smith and Rhoderick (1970).)**

| Metal | $\phi_m$ (eV) | $\phi_b^p$ (eV) | $\phi_b^n$ (eV) | $\phi_b^p + \phi_b^n$ |
|---|---|---|---|---|
| Al | 4·20 | 0·58 | 0·50 | 1·08 |
| Au | 4·70 | 0·35 | 0·81 | 1·16 |
| Cu | 4·52 | 0·46 | 0·69 | 1·15 |
| Ni | 4·74 | 0·51 | 0·67 | 1·18 |
| Pb | 4·20 | 0·55 | 0·41 | 0·96 |

## 4. I–V characteristics

There are three principal conduction mechanisms in a Schottky junction. For the case of an n-type semiconductor these are:

(i) transport of electrons from the conduction band of the semiconductor over the top of the barrier into the metal,

(ii) injection of holes into the neutral region of the semiconductor, and

(iii) recombination (or generation) of electrons and holes in the depletion region of the semiconductor.

These processes are shown schematically in figure 6. Since (ii) involves hole-electron recombination, it differs from (iii) only in that the recombination takes place in two different regions. There is also in some circumstances a fourth process, namely tunnelling of elec-

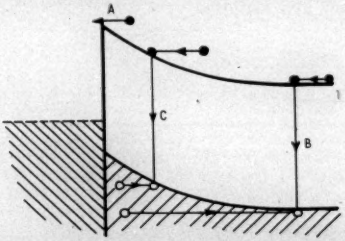



Figure 6. Schottky barrier, showing conduction mechanisms under forward bias. A, emission over barrier; B, recombination in neutral region (hole injection); C, recombination in space-charge region.

trons from the conduction band into the metal. This was at first thought to be the most important transport mechanism until it was realized that it predicted the wrong direction of rectification. It is now known to play a very minor role in the majority of practically important cases.

It is possible to make metal–semiconductor junctions in which processes (ii) and (iii) are insignificant in comparison with (i). Such junctions are said to be 'ideal', and we may think of (ii) and (iii) as causing small departures from ideality.

### 4.1. *Transport of electrons over the barrier*

Two apparently quite distinct theories have been proposed to describe process 4(i). Historically, the first of these was the so-called diffusion theory of Wagner (1931) and Schottky and Spenke (1939). According to this theory, the current is supposed to be limited by the ordinary processes of diffusion and drift operating within the space-charge region where the bands are bent. But according to the thermionic emission theory, proposed by Bethe (1942), the current is assumed to be limited by the emission of electrons over

the barrier into the metal in a manner similar to the thermionic emission of electrons from a metal into a vacuum. At first sight, the condition for the validity of the thermionic emission theory would seem to be that the mean free path of the electrons should be greater than the thickness of the barrier region, but a more refined argument given by Bethe suggests that all that is necessary is that the mean free path should be greater than the distance in which the barrier falls by an amount $kT$ ftom its maximum value.

The difference between the two theories can be more clearly understood by considering the behaviour of the quasi-Fermi level for electrons. The quasi-Fermi level is a hypothetical energy level which has the significance that, if inserted into the Fermi–Dirac distribution function, it gives the correct concentration of electrons even though the system is not in true thermal equilibrium. Far from the junction the quasi-Fermi level must coincide with the Fermi levels in metal and semiconductor respectively. The assumption tacitly made in the diffusion theory is that the quasi-Fermi level coincides with the Fermi level in the metal *at the junction* (figure 7). This is just another way of saying that the conduction electrons

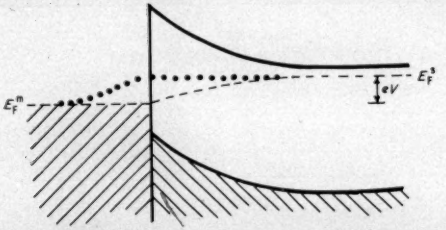



Figure 7. Schottky barrier with forward bias voltage $V$, showing behaviour of quasi-Fermi level – – – – diffusion theory; • • • • thermionic emission theory.

in the semiconductor immediately adjacent to the metal are in equilibrium with those in the metal, i.e. that the concentration of electrons in the semiconductor at the top of the barrier does not change when a bias is applied. Since the gradient of the quasi-Fermi level is the driving force for electrons, this theory implies that the bottleneck for current flow is provided by the transport of electrons in the space charge region.

The assumption that the quasi-Fermi level is coincident with the Fermi level in the metal at the junction is in sharp contrast with the assumption normally made in p–n junction theory, where it is assumed that the quasi-Fermi level for each type of carrier remains flat through the junction. This assumption has been justified at length by Shockley (1949). Since the transport mechanism in a Schottky barrier is basically very similar to that in a p–n junction, there seems to be a contradiction here. The contradiction was resolved by Gossick (1963), who pointed out that the electrons which cross the barrier from the semiconductor into the metal are not in thermal equilibrium with the conduction electrons in the metal because they have energies exceeding the Fermi-energy by about 1 eV (see figure 8). In a loose sense, they are 'hot' electrons, and in a crude sort of way we may regard them as a different species of electrons from the ordinary 'cool' electrons in the metal, having a quasi-Fermi level which differs from the Fermi-level in the metal. As the hot electrons move into the metal, they lose energy by collisions, and their quasi-Fermi level falls to that of the metal (figure 7). This allows us to relax the requirement that the quasi-Fermi level at the boundary must coincide with the Fermi level in the metal; in particular, we may now go to the other extreme and assume that the quasi-Fermi level remains constant throughout the barrier region, as in a p–n junction. This is precisely the approximation made in the thermionic emission theory. It is equivalent to assuming that the concentration of electrons at the top of the barrier changes by a factor $\exp(eV/kT)$ when a forward bias voltage $V$ is applied.

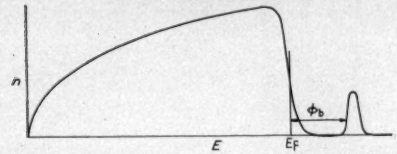



Figure 8. Energy distribution of electrons in metal. The parabolic distribution falling off at $E_F$ represents the ordinary conduction electrons. The small 'pip' (much exaggerated) at an energy $E_F+\phi_b$ represents the 'hot' electrons injected into the metal from the semiconductor.

It can be shown by analysing existing experimental data (E. H. Rhoderick, to be published) that for high mobility semiconductors (silicon, gallium arsenide etc.) the thermionic emission theory is a good approximation and the diffusion theory a very bad one In other words, the bottleneck for current flow is provided by the actual process of emission of electrons into the metal rather than by the transport of electrons through the space charge region. According to the thermionic emission theory, the current density for an applied voltage $V$ is given (see Henisch 1957) by

$$J=J_0\{\exp(eV/kT)-1\} \tag{3}$$

where

$$J_0=AT^2\exp(-\phi_b/kT). \tag{4}$$

Here $A$ is the Richardson constant modified for the effective mass of the electrons in the semiconductor (Crowell 1965). For moderately large forward bias voltages such that $V>3\,kT/e$, the second term in the expression for $J$ is negligible, and we may write

$$J=J_0\exp(eV/kT). \tag{5}$$

Junctions whose $J$–$V$ characteristics can be represented by (3) or (5) are said to have 'ideal' characteristics.

### 4.2. *Hole injection*

The position in the literature with regard to hole injection is confusing. The early papers on the subject (e.g. Bardeen and Brattain 1949, Banbury 1953), which were all concerned with point contacts on n-type germanium, predicted that in the forward direction the current should be largely carried by holes, with an injection ratio approaching unity. This conclusion seems to have been based on the purely qualitative argument that, if the barrier height exceeds one-half of the energy gap, the surface of the semiconductor immediately underneath the metal is inverted, so that at the interface the density of holes exceeds the density of electrons. But the downward bending of the bands away from the interface presents a barrier to holes, and the injection ratio is determined not by the ratio of the electron and hole densities at the interface but by the ratio of the electron density at the interface to the hole density in the bulk of the semiconductor. This is generally much greater than unity even though the surface may be inverted.

Scharfetter (1965) has calculated the injection ratio in Schottky diodes by using standard p–n junction theory to calculate the flow of holes in the neutral region and the thermionic emission theory to calculate the electron current. He finds that for small forward bias voltages, for which the holes flow by diffusion, the injection ratio is very small (of the order of $10^{-4}$ in a typical case), but that above a critical current density ($\sim 10^4$ A m$^{-2}$) the electric field in the semiconductor assists the flow of holes and the injection ratio increases linearly with current density. Scharfetter's results have been confirmed experimentally by Yu and Snow (1969) for diodes made by evaporating gold films onto silicon.

It now seems firmly established, both experimentally and theoretically, that for large-area evaporated-film Schottky diodes on n-type semiconductors the injection ratio for holes is exceedingly small at low current densities. This makes it hard to understand how a point-contact transistor works; the probable explanation is that in a point-contact emitter the current densities are quite high ($\sim 10^5$ A m$^{-2}$), and exceed the critical value at which the injection ratio begins to increase.

### 4.3. *Recombination or generation in the depletion region*

Unlike hole injection, recombination (for forward bias) or generation (for reverse bias) within the depletion region may be very important at comparatively low current levels. The role of these processes was convincingly demonstrated in a classic paper by Yu and Snow (1968), who added a gate electrode which covered the surface of the semiconductor (in their case n-type silicon) immediately adjacent to the metal contact, as shown in figure 9. This enabled them to alter the potential of the semiconductor surface as in a MOS transistor, and to alter the area of the depletion region at will. By putting a negative bias on the gate, the surface of the semiconductor surrounding the metal contact could be strongly inverted, thereby increasing the area of the depletion region (see figure 10(*b*)). Because the p-type skin induced on the surface was in good electrical contact with the metal, this effectively increased the area for process 4(iii) without changing the area for process 4(i). In this way

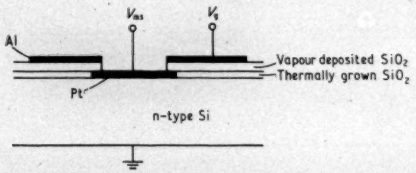



Figure 9. Schottky barrier with gate electrode (Yu and Snow 1968).

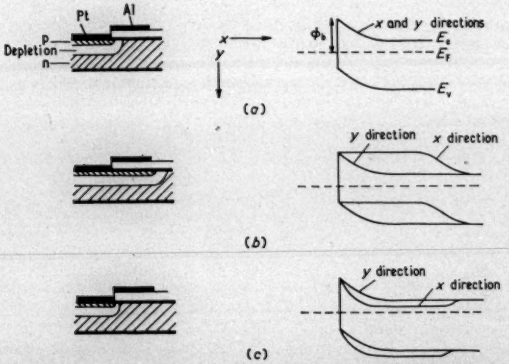



Figure 10. Band diagrams for Schottky barrier with gate (Yu and Snow 1968). (*a*) Flat bands ($V_G = V_{FB}$), (*b*) Surface inverted ($V_G \ll V_{FB}$), (*c*) Surface accumulated ($V_G \gg V_{FB}$).

the two processes could be separated, and it was concluded that for moderately large forward bias voltages the current is of the form

$$J=J_0 \exp(eV/kT)+J_r \exp(eV/2kT)$$

where the first term is the ordinary thermionic emission current given by (5), and the second term is the recombination current. $J_r$ is given by

$$J_r=en_i dA'/\tau,$$

where $n_i$ is the intrinsic electron concentration which is proportional to $\exp(-E_g/2kT)$, $d$ is the thickness of the depletion region, $A'$ its area, and $\tau$ the lifetime within the depletion region. If the negative bias on the gate is reduced until the bands adjacent to the metal contact are flat, the area of the depletion region $A'$ becomes equal to the area of the contact, and the recombination current is reduced in proportion (figure 10(*a*)).

For a given semiconductor, the relative importance of thermionic emission and recombination/generation current may vary from one junction to another. For flat band conditions ($A=A'$), the ratio of the two terms depends on $\phi_b$, $E_g$, $T$, and $\tau$. It is possible in practice to find combinations of these such that the 'ideal' thermionic emission current dominates, but recombination/generation current represents a very common cause of departure from the ideal thermionic emission theory. If recombination current is important, the temperature variation of the forward current shows two activation energies (figure 11). Above room temperature the activation energy tends to the value $\phi_b$, characteristic of the thermionic emission term, while below room temperature it approaches the value $E_g/2$ characteristic of the recombination current.

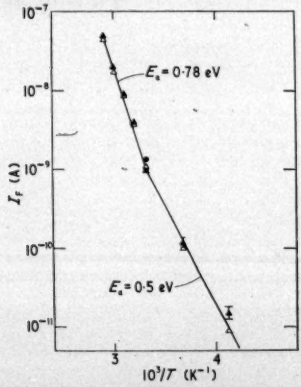



Figure 11. Forward current of platinum–silicon Schottky diode as function of temperature, for flat-band conditions. (Yu and Snow 1968.)

The reverse current under conditions where the surface adjacent to the metal is inverted or exhibits flat bands can be explained in terms of generation current. The current density can be well represented by

$$J=J_0+J_g$$

where the generation current is proportional to (reverse bias)$^{-1/2}$ because of the increase in thickness of the depletion region, and is also proportional to $A'$. The temperature dependence again shows two activation energies.

4.4. *The effects of image force lowering of the barrier and of funnelling*

Any charged body close to a conducting plane produces an electrostatic field as if there were an equal and opposite charge located at the mirror image of the body in the plane separating the conducting and insulating media. This image charge, being of opposite sign, lowers the potential energy of the body. The effect on a Schottky barrier is as shown in figure 12, and causes a lowering of the barrier by an amount $\Delta\phi$. The barrier lowering

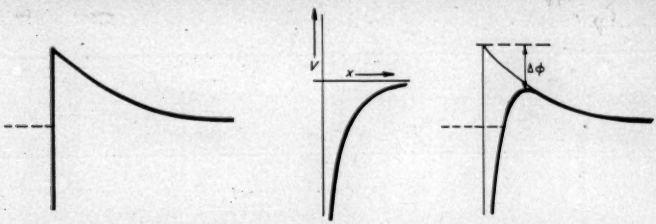



Figure 12. Image-force lowering of barrier. (*a*) Schottky barrier, (*b*) Image potential, (*c*) Resultant.

$\Delta\phi$ depends on the maximum electric field in the barrier (i.e. the field which would exist if there were no image force) and is therefore a function of the applied bias. The effects of this are twofold: (i) In the forward direction, the barrier increases with increasing forward bias so that $J$ increases less rapidly with $V$ than it would otherwise do. This changes the voltage-dependence of the forward characteristic to

$$J=J_0\{\exp(eV/nkT)-1\}$$

where $n$ is a constant slightly larger than unity—typically 1·02 for gold on 1 Ω-cm silicon (Bethe 1942). (ii) In the reverse direction, the barrier decreases with increasing reverse bias, so that $J$ does not saturate at the value $J_0$ given by (4). Instead, $J$ increases gradually with reverse bias according to $J=J_0\exp(e\Delta V/kT)$, where

$$\Delta V=\{e^3N_d(V_d-V)/8\pi^2\epsilon_0^3\epsilon_s^3\}^{1/4}$$

is the lowering of the barrier due to the image force (Sze *et al.* 1964). Here $N_d$ is the donor concentration and $\epsilon_s$ the relative permittivity of the semiconductor.

The effect of quantum-mechanical tunnelling through the barrier is important only for very thin barriers. For moderately doped semiconductors (up to $N_d\sim10^{16}$ cm$^{-3}$ say) as are normally used in practical devices, tunnelling is significant only in the reverse direction where, although the width of the depletion region increases with reverse bias, the width of that part of the barrier through which electrons have to tunnel decreases. The effect of this is to make the reverse current increase rapidly above a certain voltage, as in a Zener diode.

Both image-force lowering and tunnelling may be exaggerated if the surface of the semiconductor outside the metal contact is accumulated. In this case the depletion layer is much thinner in a direction parallel to the surface than it is in a direction normal to the surface (see figure 10(*c*)), with the result that tunnelling and image-force lowering are both enhanced. In consequence, the '*n*' value in the forward direction is a function of voltage (Yu and Snow 1968), and in the reverse direction the current increases very quickly with reverse bias, as shown in figure 13. Here the slow increase up to $V=-0{\cdot}3$ V is due to image-force lowering and the fast increase above $V=-1{\cdot}0$ V is due to tunnelling.

The 'ideality' of the $I$–$V$ characteristics (low $n$ value, good saturation in the reverse direction, high reverse breakdown) is thus very sensitive to the semiconductor surface potential

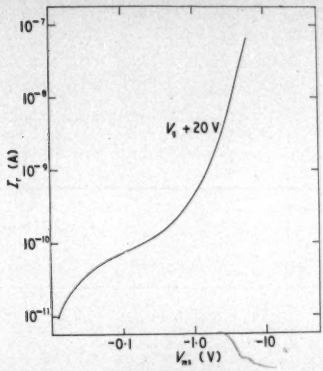



Figure 13. Reverse *I–V* characteristic for Pt–Si Schottky diode with accumulated surface (Yu and Snow 1968).

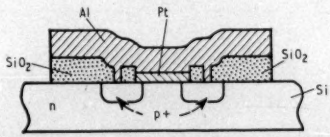



Figure 14. Schottky diode with p-type guard ring (Lepselter and Sze 1968).

adjacent to the metal, and in practice departures from ideality can often be associated with variations in surface potential resulting from surface films, adsorbed atoms or ions, etc. These effects can be largely overcome by using a guard ring consisting of a diffused p-type annulus (Lepselter and Sze 1968) as shown in figure 14. This has the effect of ensuring that the surface of the semiconductor immediately adjacent to the metal is inverted, thereby avoiding any pronounced effects due to image-force lowering or tunnelling. There will, of course, be an increase in recombination–generation current due to an increase in the area of the depletion layer, but this effect is relatively small. The characteristics of diodes prepared in this way are close to ideal.

### 4.5. *The effect of an insulating interfacial layer*

In practice, a metal-semiconductor junction, unless it is made by cleavage in ultra-high vacuum, will inevitably include a thin insulating layer of oxide between metal and semiconductor. This film is normally so thin (~10 Å) that electrons can tunnel though it quite freely. However, when a voltage difference exists across the junction, there will be a voltage drop across the film with the result that the change in the barrier height as seen from the semiconductor is less than it would otherwise have been. This makes the forward characteristic increase more slowly than $\exp(eV/kT)$. If the interfacial layer is so thin that the occupation of the surface states is determined by the Fermi level in the metal, it can be shown (Crowell and Sze 1966, Smith 1969) that the forward current is approximately proportional to $\exp(eV/nkt)$, where

$$n = 1 + \{\delta\epsilon_s\epsilon_0/\lambda(\epsilon_1\epsilon_0 + e\delta D_s)\},$$

$\epsilon_i$ being the relative permittivity of the insulating film, $\epsilon_s$ the relative permittivity of the semiconductor, and $\lambda$ the width of the depletion layer. Since $\lambda$ is proportional to $(V_d - V)^{1/2}$, $n$ is not strictly constant but varies slightly with forward bias.

### 4.6. *Causes of departure from ideal characteristics*

To sum up, it is possible to make Schottky junctions in which the thermionic emission of electrons over the barrier is far more important than any other current mechanism. Such diodes have $J$–$V$ characteristics which approximate very closely to (3). Deviations from this ideal $J$–$V$ relationship may be due to:

(i) Recombination–generation current (composite forward characteristic with two activation energies; non-saturating reverse current).

(ii) Image force lowering of the barrier ($n$ value slightly greater than 1; non-saturating reverse current).

(iii) Tunnelling (important only in reverse direction for moderate doping).

(iv) Insulating interfacial layer ($n$ value greater than 1).

The first three of these are sensitive to the potential of the surface of the semiconductor immediately adjacent to the metal. If surface effects can be minimized, either by using a guard ring or by great attention to surface cleanliness, near-ideal characteristics can be obtained. For example, Lepselter and Sze (1968) obtained an exponential $J$–$V$ relationship over 8 orders of magnitude and an $n$ value of 1·02 for Pt–Si diodes using a p-type guard ring, and Smith (1968) has obtained similar near-ideal behaviour for Au–GaAs diodes, but with an $n$ value of 1·06, by using a special cleaning procedure.

## 5. Transient effects

### 5.1. *Recovery time*

The recovery time which determines the switching speed of a Schottky diode depends on storage times of carriers in various parts of the device. Although the Schottky diode is commonly said to be a 'majority carrier' device, this is not strictly true. The electrons injected into the metal ('hot' electrons) are effectively minority carriers in the metal and can be sucked back by the application of reverse bias until they lose sufficient energy to stop them returning over the barrier. Roughly speaking, this happens when an injected electron collides with a conduction electron within the metal. The mean free path for this process is about 500 Å and the velocity about $10^6$ m $s^{-1}$, so the storage time for injected electrons is about $10^{-13}$ s or less, which is negligible for most purposes. The storage time for injected holes is longer, and is determined by the recombination time which is typically $10^{-6}$ s for silicon. However, the contribution of holes to the recovery time is much smaller than this due to the very low injection ratio. The stored charge due to holes is $I_p\tau_p$, where $I_p$ is the hole contribution to the forward current and $\tau_p$ the hole lifetime. If the circuit parameters are such as to limit the reverse current after switching to $I_r$, then the effective recovery time is $I_p\tau_p/I_r$. If $I_r$ is of the order of the forward current, which is predominantly made up of the electron component $I_n$, then the effective recovery time is of the order of $I_p\tau_p/I_n$, which is reduced by the injection ratio $I_p/I_n$. In practice, recovery times of Schottky diodes are limited by their RC product rather than by charge storage, and times shorter than $10^{-11}$ s have been recorded for silicon diodes.

### 5.2. *Bulk traps*

It is not uncommon for some semiconductors (especially gallium arsenide) to contain deep-lying traps around the middle of the gap, and in gallium arsenide these can often have time constants of the order of seconds or minutes. These bulk traps may give rise to transient effects. For example, if the junction is reverse biased, the quasi-Fermi level for electrons may fall below the trap level and the traps begin to empty and become positively charged. The effect is as if the donor density were increased from $N_d$ to $N_t + N_d$, where $N_t$ is the trap density. The depletion layer gets thinner and the capacitance increases with

the time constant of the traps. If one waits until the capacitance is steady before making the measurement, the $C^{-2}$–$V$ plot will have two straight line portions which intersect when the quasi-Fermi level for electrons first falls below the trap energy $E_t$. This gives a means of measuring $E_t$. The effect may be very undesirable if the $C^{-2}$–$V$ plot is being used to measure $N_d$. The difficulty may be overcome if the diode is biased in the forward direction immediately before measurement and the measurement then made before the traps have time to empty. In this way true values of $N_d$ may be obtained even in the presence of traps (Smith and Rhoderick 1969).

## 6. Conclusion

The present position (1970) seems to be that the basic properties of Schottky barriers are fairly well understood. There are still some areas (e.g. precise dependence of barrier height on work-function of metal, exact causes of departure from ideal $J$–$V$ characteristics) where the agreement between theory and experiment is imperfect, but in most cases this is due to inadequate knowledge of the nature of the interface or of the surface properties of the semiconductor, and stems from limitations in the technology of fabrication rather than deficiencies in the theory. It is probably fair to say that Schottky junctions are now better understood than p–n junctions.

## References


ARCHER, R. J., and ATALLA, M. M., 1963, *Ann. N.Y. Acad. Sci.*, **101**, 697–708.

ATALLA, M. M., 1966, *Proc. Munich Symp. on Microelectronics* (Munich: Oldenburg), pp. 123–57

BANBURY, P. C., 1953, *Proc. Phys Soc.* B, **66**, 833–40.

BARDEEN, J., 1947, *Phys. Rev.* **71**, 717–27.

BARDEEN, J., and BRATTAIN, W. H., 1949, *Phys. Rev.*, **75**, 1208–25.

BETHE, H. A., 1942, *M.I.T. Radiat. Lab. Rep.*, 43–12.

BRAUN, F., 1874, *Ann. Phys., Lpz.*, **153**, 556–62.

COWLEY, A. M., and SZE, S. M., 1965, *J. appl Phys.*, **36**, 3212–20.

CROWELL, C. R., 1965, *Solid-St. Electron.*, **8**, 395–9.

CROWELL, C. R., and SZE, S. M., 1966, *Solid-St. Electron.*, **9**, 1035–48.

GOSSICK, B. R., 1963, *Solid-St. Electron.*, **6**, 445–52.

HENISCH, H. K., 1957, *Rectifying Semiconductor Contacts* (Oxford: Clarendon Press).

JONES, R. O., 1968, *Phys. Rev. Lett.*, **20**, 992–4.

LEPSELTER, M. P., and SZE, S. M., 1968, *Bell Syst. Tech. J.*, **47**, 195–208.

MEAD, C. A., 1966, *Solid-St. Electron.*, **9**, 1023–33.

MEAD, C. A., and SPITZER, W. G., 1964, *Phys. Rev.*, **134**, A713–16.

PICKARD, G. W., 1906, U.S. Patent No. 836531.

PIERCE, G. W., 1907, *Phys. Rev.*, **25**, 31–63.

POOLE, K. F., 1969, *Ph.D. Thesis*, University of Manchester.

SCHARFETTER, D. L., 1965, *Solid-St. Electron.*, **8**, 299–311.

SCHOTTKY, W., 1938, *Naturwiss.*, **26**, 843–53.

SCHOTTKY, W., and SPENKE, E., 1939, *Wiss. Veroff. Siemens Werke*, **18**, 1–67.

SHOCKLEY, W., 1949, *Bell Syst. Tech. J.*, **28**, 435–89.

SMITH, B. L., 1969, *Ph.D. Thesis*, University of Manchester.

—— 1968, *Electron. Lett.*, **4**, 332–3.

SMITH, B. L., and RHODERICK, E. H., 1969, *J. Phys.* D: *Appl. Phys.*, **2**, 465–7.

—— 1970, *Solid-St. Electron.*, in the press.

SZE, S. M., 1969, *Physics of Semiconductor Devices* (New York: John Wiley).

SZE, S. M., CROWELL, C. R., and KAHNG, D., 1964, *J. appl. Phys.*, **35**, 2534–6.

TURNER, M. J., and RHODERICK, E. H., 1968, *Solid-St. Electron.*, **11**, 291–300.

VAN LAAR, J., and SCHEER, J. J., 1967, *Surf. Sci.*, **8**, 342–56.

WAGNER, C., 1931, *Phys. Z.*, **32**, 641–5.

YU, A. Y. C., and SNOW, E. H., 1968, *J. appl. Phys.*, **39**, 3008–16.

—— 1969, *Solid-St. Electron.*, **12**, 155–60.

J. W. POWELL
UKAEA, Winfrith, Dorchester, Dorset

# A review of progress in gas lubrication

**Abstract.** The past decade has witnessed a rapid evolution in the technology of gas lubrication which has been accompanied by the application of gas bearings on an increasing scale. Before 1960, existing applications were restricted to two specialized areas: gyroscopes for inertial navigation and gas circulators in nuclear reactors. Now they are widely employed in manufacturing industry, in grinding machines, drilling machines and hand tools and in a wide range of medical and scientific instruments. This progress has been made possible by the solution of two or three major technical problems and the assimilation of myriad minor advances in structural design, materials and manufacturing techniques.

This review describes the theory and practice of journal and thrust bearings of the self-acting and externally-pressurized types. Self-acting bearings depend upon rotation for the generation of load supporting pressure by the mechanism of viscous shearing. The influence of gas compressibility on the steady-state load capacity of journal bearings is described in some detail. The problem of self-excited half-speed whirl is discussed together with the several methods employed for improving bearing stability. Self-acting thrust bearings of the spirally grooved, stepped sector and pivoted pad types are described.

Externally-pressurized bearings depend for their operation on a continuous throughflow of gas from a pressurized source. Journal bearings can support a load at zero speed but the load capacity is enhanced by rotation in a similar manner to self-acting bearings. The theory of both steady-state and dynamic performance is reviewed and some typical applications are described. The theory and application of externally-pressurized thrust bearings is also discussed.

## 1. Introduction

At the present time most rotating machines are lubricated by liquid or solid lubricants in the familiar form of oils and greases. In the majority of applications these conventional lubricants offer an acceptable standard of efficiency and reliability. However, as technology progresses applications are continuously arising where properties are required which are not possessed by any known liquid or solid lubricant. In some such circumstances the answer has been found in the use of gas-lubricated bearings.

The most distinctive property of gaseous lubricants when compared to oils or greases is their very low viscosity. The viscosity of air is approximately ten thousand times less than that of a light oil at ambient temperature. There are many applications where low viscosity is desirable at both low and high rotational speeds. Some precision instruments such as dynamometers, balances, viscometers and roundness measuring machines require bearings which are free from stick-slip motion at low speed and in these applications externally-pressurized air bearings are becoming firmly established. At high speeds the advantages of low viscosity are equally apparent. Frictional heating is reduced and as a result not only is the efficiency of the machine increased but problems of thermal expansion, distortion of the bearing surfaces and changes of lubricant properties are often avoided. However, in many high-speed applications of gas lubricated bearings the reduction in friction has been of only secondary importance. Most conventional high-speed machines employ rolling contract bearings in which balls or rollers run between inner and outer races. These bearings require oil or grease lubrication and though they operate satisfactorily in the majority of applications there are some circumstances under which their performance is unsatisfactory either owing to the vibration and noise which they generate

or through some limitation of their lubricant. Most widescale applications of gas-lubricated bearings have arisen because of one or a combination of the problems arising with rolling contact bearings, and the two earliest of these applications did much to stimulate research and development.

The introduction of nuclear submarines capable of remaining submerged for several months without the opportunity to make star sightings demanded an inertial navigation system greatly superior to that available hitherto. The accuracy of an inertial navigation system depends upon the accuracy of the direction sensing element which is conventionally a gyro wheel mounted on precision bearings. The performance of the gyro is dependent upon the quality of its bearings since spurious precessions known as 'wander' result from unwanted torques arising from friction and vibration. In such an application ball bearings not only limit performance from the outset but gradually deteriorate in service to an extent that their useful life is sometimes as little as a thousand hours. On the other hand self-acting gas spin axis bearings have been developed with low friction, an undetectable vibration level and no change in properties with time. This development greatly stimulated interest in gas lubrication in the USA in the middle 1950s, and research in the UK began at the Admiralty Compass Observatory, Slough, in 1959.

The field of nuclear engineering provided the other forcing ground for gas bearing research. There are many instances in gas-cooled power reactors, in fuel processing plant and in experimental loops where it is convenient to locate gas pumps and circulators entirely within the gas circuit. By this means rotating shaft seals are eliminated and the risk of leakage of radio-active material is minimized. However, since the bearings of the gas circulators must then operate in the radio-active environment, conventional lubricants cannot be used because they deteriorate rapidly through polymerization. The most satisfactory answer appeared to be to use the process gas as the bearing lubricant thereby avoiding both deterioration of the lubricant and contamination of the gas circuit. However, before this was achieved much research and development needed to be undertaken and some interesting problems were solved. This work was carried out in the UK at AERE Harwell in the middle 1950s.

The past decade has witnessed a rapid evolution in the technology of gas lubrication which has been accompanied by the application of gas bearings on an increasing scale. Before 1960, applications of gas bearings were restricted to the two specialized areas already mentioned. Now they are widely employed in manufacturing industry in grinding machines, drilling machines and hand tools and in a wide range of medical and scientific instruments. This progress has been made possible by the solution of two or three major technical problems and the assimilation of myriad minor advances in structural design, materials and manufacturing techniques.

### 1.1. *Types of gas bearings*

It has become customary to classify gas bearings according to the method employed in generating the load-carrying pressure in the gas film between the bearing surfaces. The three principal types are self-acting or aerodynamic, externally pressurized or aerostatic and squeeze film.

Self-acting bearings derive their load-carrying capacity from the rotation of one of the bearing surfaces. Pressure is generated by the viscous shearing of the gas film and the mechanism is similar to that occurring in hydrodynamic oil bearings. Neglecting the effect of compressibility, both the friction and load capacity of self-acting bearings is proportional to the lubricant viscosity. Thus for bearings of the same size and clearance, and operating at the same speed, the load capacity and friction of an air bearing is some four orders of magnitude lower than that of the corresponding bearing lubricated by SAE 10 oil. Load capacity increases with speed and at high speeds self-acting gas bearings can carry useful loads. However, specific loadings are always very much lower than for oil-lubricated bearings, and the absence of any boundary lubrication effect allows the bearing

surfaces to rub at low speeds during starting and stopping. This has been found to cause rapid wear in bearings with specific loadings higher than 2 lbf $in^{-2}$ (140 g $cm^{-2}$) if metallic bearing surfaces are used. For this reason self-acting gas bearings are most suitable for machines that run for long periods at constant speed and start and stop infrequently. They have found application in spin axis bearings for gyroscopes and in gas circulators and compressors. However as they require more precise manufacture than externally pressurized bearings and because they are somewhat less robust, they have not been applied in any large-scale industrial applications. There have been attempts to develop self-acting air bearings for small electrically-driven fans and for textile spindles but these have so far been unsuccessful due to high manufacturing costs.

Externally pressurized bearings are supplied with gas under pressure from an external source that is usually a compressor. There is a continual flow of gas through the bearing and hence an expenditure of power in maintaining the lubricant film. The load capacity of externally pressurized bearings is proportional to the gas supply pressure and is independent of the gas viscosity. Thus externally pressurized bearings can carry much higher loads than self-acting bearings and their load capacity is not necessarily dependent upon rotation. However aerodynamic effects can increase the load capacity of externally pressurized journal bearings at high rotational speeds, and bearings designed to take advantage of this augmentation are often called hybrid bearings. To date all widescale industrial applications of gas lubrication have involved externally pressurized air bearings and all such bearings operating at high speeds are hybrid in operation. Externally pressurized air bearings are used in the wheelheads and workheads of precision grinding machines, in high-speed internal-grinding spindles, in high-speed drilling spindles for electronic printed circuit boards and microhole drilling in metals and in precision-boring machines. They have also been applied in industrial hand tools, dentists' drills, orthopaedic drills, gas flowmeters and a variety of scientific and industrial instruments.

In squeeze film bearings one of the surfaces is driven to vibrate in the direction normal to the plane of the bearing surfaces so that the gas film thickness is caused to fluctuate. The pumping effect in the gas film generates pressure which can support a load. There is no dependence on an external gas supply or upon rotation but the problems of development of an efficient vibration generator has retarded applications and no widescale use has yet emerged. For this reason the present work will be confined to a review of research into self-acting and externally pressurized bearings only.

## 2. Self-acting gas journal bearings

### 2.1. *Steady-state conditions*

Figure 1 shows a simple slider bearing in which one surface is stationary and the other surface moves with constant velocity $u$. In the analysis of the behaviour of the fluid between the bearing surfaces the following assumptions are made.

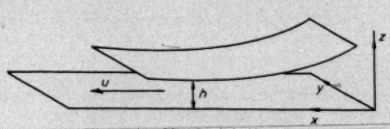



Figure 1. Simple slider bearing.

(i) The flow is laminar.
(ii) The fluid behaves as a Newtonian fluid.
(iii) There is no slip at the boundaries between the surfaces and the fluid.
(iv) Inertia effects are negligible with respect to pressure gradient effects.
(v) Pressure is constant across the fluid film perpendicular to the bearing surfaces.

(vi) Fluid velocities perpendicular to the bearing surfaces may be neglected with respect to velocities parallel to the bearing surfaces.
(vii) Fluid viscosity is constant across the fluid film.

When these assumptions are applied the Navier Stokes equations reduce to

$$\frac{\partial P}{\partial x} = \mu \frac{\partial^2 u}{\partial z^2}$$

$$\frac{\partial P}{\partial y} = \mu \frac{\partial^2 v}{\partial z^2}$$

where $P$ is the absolute pressure at any point $(x, y, z)$ and $u$ and $v$ are the velocities in the $x$ and $y$ directions respectively, and $\mu$ is the coefficient of viscosity of the fluid.

Integrating for the boundary conditions $u = U$ at $z = 0$, $u = 0$ at $z = h$, $v = 0$ at $z = 0$ and $z = h$, and applying the continuity equation for compressible flow

$$\frac{\partial(\rho u)}{\partial x} + \frac{\partial(\rho v)}{\partial y} + \frac{\partial(\rho w)}{\partial z} = 0$$

where $\rho$ is the fluid density and $w$ is the fluid velocity in the $z$ direction, finally yields:

$$\frac{\partial}{\partial x}\left(\rho h^3 \frac{\partial P}{\partial x}\right) + \frac{\partial}{\partial y}\left(\rho h^3 \frac{\partial P}{\partial y}\right) = 6\mu U \frac{\partial}{\partial x}(\rho h). \tag{1}$$

Equation (1) with constant density reduces to the familiar Reynolds equation for incompressible lubrication. In order to solve equation (1) for a particular bearing it is necessary to define a relationship between the film thickness $h$ at any point and the $x$ and $y$ coordinates. Consider the journal bearing shown in figure 2 with cylindrical polar coordinates $\theta = x/a$,

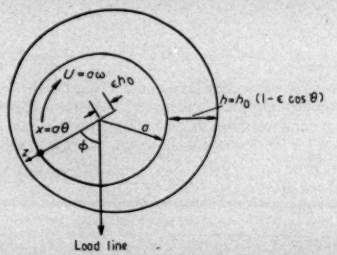



Figure 2. Coordinate system for self-acting journal bearing.

$y$ and $z$ where $a$ is the journal radius and $z$ is measured perpendicular to the journal surface. The gas film thickness $h$ is independent of $y$ and to a good approximation $h = h_0(1 - \epsilon \cos \theta)$. $h_0$ is the mean radial clearance of the bearing, $\epsilon$ is the fractional eccentricity of the journal in the bearing and $\theta = 0$ at the point of minimum film thickness. $h_0/a$ is typically less than $10^{-3}$ and so any effects of curvature can be neglected.

In the case of a compressible lubricant it is necessary to define a relationship between fluid density and pressure. In general

$$\rho = \rho_a \left(\frac{P}{P_a}\right)^{1/k}$$

where $P_a$ is the ambient pressure, $\rho_a$ is the fluid density at ambient pressure and $k$ is the index of expansion of the gas.

The index of expansion $k$ can have any value between 1 for isothermal conditions and $\gamma = C_p/C_v$ for adiabatic conditions, $C_p$ is the specific heat at constant pressure and $C_v$ is the specific heat at constant volume.

Considering a bearing of infinite length, i.e. $P$ independent of $y$, equation (1) reduces to

$$\frac{\partial}{\partial x}\left(\rho h^3 \frac{\partial P}{\partial x}\right) = 6\mu U \frac{\partial}{\partial x}(\rho h).$$

Then changing to cylindrical polar coordinates, substituting for $h$ and $\rho$ and integrating with respect to $\theta$ gives:

$$\frac{dP}{d\theta} = \frac{\Lambda P_a}{(1-\epsilon\cos\theta)^2}\left\{1 + \frac{C}{1-\epsilon\cos\theta}\left(\frac{P_a}{P}\right)^{1/k}\right\} \tag{2}$$

where $\Lambda = 6\mu a U/h_0^2 P_a$.

This equation was first derived for isothermal gas conditions ($k=1$) by Harrison (1913). Integration of equation (2) would give the pressure distribution around the gas film of a bearing of infinite length. Unfortunately no exact analytical solution of Harrison's equation has been found and it is possible that none exists. However solutions have been obtained by numerical methods and some good approximate analytical solutions have been derived.

It is interesting that exact solutions of equation (2) can be found for the limiting cases of $\Lambda = 0$ and $\Lambda = \infty$. At $\Lambda = 0$ the solution is the well-known Sommerfeld equation for incompressible lubrication

$$\frac{\bar{P}}{P_a} = \frac{\Lambda\pi\epsilon}{(2+\epsilon^2)(1-\epsilon^2)^{1/2}}$$

and

$$\phi = \frac{\pi}{2} \qquad \text{(independent of } \Lambda \text{ and } \epsilon)$$

where $\bar{P}$ is the bearing load per unit plan area and $\phi$, the attitude angle, is the angle between the load line and the point of minimum clearance as shown in figure 2.

At $\Lambda = \infty$ the effect of compressibility is overriding and the load capacity is independent of $\Lambda$. The solution is however dependent upon the index of expansion of the gas. For isothermal conditions

$$\frac{\bar{P}}{P_a} = \frac{\pi\{1-(1-\epsilon^2)^{1/2}\}}{\epsilon(1-\epsilon^2)^{1/2}} \tag{3}$$

and for adiabatic conditions

$$\frac{\bar{P}}{P_a} = \frac{\pi\epsilon\gamma}{2}\left\{1 + (\gamma+1)(\gamma+2)\frac{\epsilon^2}{8} + (\gamma+1)(\gamma+2)(\gamma+3)(\gamma+4)\frac{\epsilon^4}{192}\right\} \tag{4}$$

for both solutions $\phi = 0$.

$\Lambda$ is called the compressibility number as its magnitude for a particular bearing indicates the significance of compressibility on the bearing performance. The effect of compressibility upon the form of the circumferential pressure distribution is shown in figure 3. At low compressibility numbers the pressure distribution is symmetrical disposed above and below the ambient pressure. At higher compressibility numbers the pressure distribution becomes increasingly assymetric with the positive pressure region greatly accentuated. It has become customary following Ford *et al.* (1957) to present journal bearing load carrying performance on dimensionless graphs of $\bar{P}/P_a$ versus $\Lambda$ for constant eccentricity. Figure 4 shows one such graph and includes the assymptotic solutions at $\Lambda = \infty$ for isothermal and adiabatic gas conditions.

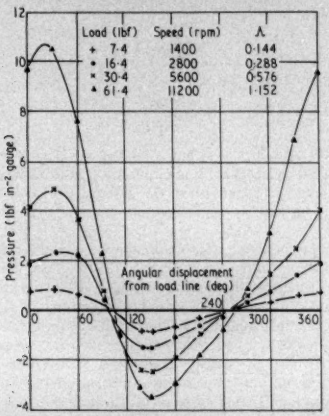



Figure 3. Experimental pressure distributions around the centre of a self-acting air journal bearing. Length 4 in, diameter 2 in, $h_0 = 1{\cdot}045 \times 10^{-3}$ in, $\epsilon = 0{\cdot}5$.

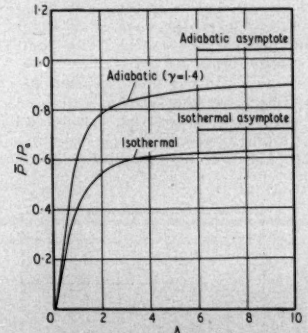



Figure 4. Theoretical effect of gas behaviour on journal bearing load capacity ($\epsilon = 0{\cdot}4$)—Infinitely long bearing (after Ausman 1957 a).

The curves shown on figure 4 are derived from the first order perturbation solution of equation (2) which was first derived by Ausman (1957 a). Ausman expressed the gas film pressure in a series of ascending powers of eccentricity ratio:

$$P = P_a + \epsilon P_1 + \epsilon^2 P_2 + \epsilon^3 P_3 \ldots .$$

To second order terms in $\epsilon$ it is shown that

$$\frac{\bar{P}}{P_a} = \frac{\pi k \epsilon \Lambda}{2(1+\Lambda^2)^{1/2}} \tag{5}$$

and

$$\phi = \tan^{-1}\left(\frac{1}{\Lambda}\right)$$

where

$$\Lambda = \frac{6\mu a U}{k h_0^2 P_a}.$$

This solution is valid for all values of $\Lambda$ but is limited to values of $\epsilon$ less than 0·5 since the higher order terms in $\epsilon$ are neglected. It is of interest that the load capacity of the bearing is proportional to the index of expansion of the gas film. Thus if the gas film conditions were adiabatic the load capacity would be proportional to the ratio of specific heats of the gas lubricant.

The earliest experiments measuring the load capacity of self-acting gas journal bearings were reported by Ford *et al.* (1957) and Wildman (1957). Ford carried out tests with air, helium, hydrogen and di-methyl-ether and remarked upon a tendency for load capacity to increase with $\gamma$. Ausman (1957 b) improved his earlier theory to take account of end leakage in finite length bearings and showed that both Ford's and Wildman's data showed a closer agreement with his adiabatic predictions than with his isothermal predictions. Similar conclusions were also reached by Katto and Soda (1961). However it seemed highly unlikely that any significant temperature variations could exist around the extremely thin gas film bounded by highly conductive metal surfaces one of which was rotating and providing a substantial thermal mixing effect. Thus many argued that in spite of experimental evidence to the contrary the gas film behaviour must be very close to isothermal.

The greatest experimental difficulty concerned the measurement of eccentricity. Most early experimental bearings were 2 in in diameter and employed radial clearances of around $10^{-3}$ in (0·025 mm) and so the measurement of eccentricity to an accuracy of 2% demanded equipment discriminating down to $2 \times 10^{-5}$ in (0·0005 mm). Such equipment was probably not available before the introduction of the Wayne Kerr capacitance transducer system in 1960. Since that time several experimenters have reported load test data in close agreement with digital computer solutions of Harrison's equation for isothermal gas conditions. Powell (1963) reported load test data for self-acting journal bearings operating with air, argon, hydrogen, helium, propane and carbon dioxide. No variation with $\gamma$ was observed and in all cases the agreement with isothermal theory was good

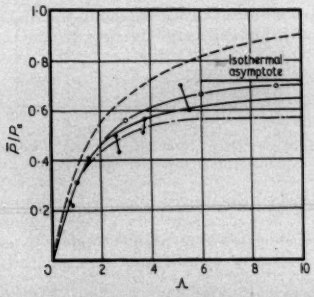



Figure 5. Load capacity of finite length gas journal bearings. Length-to-diameter ratio = 2, $\epsilon$ = 0·4. Experimental: – – – Ford *et al.* 1957; —— – Whitley and Betts 1959; — Whitley *et al.* 1962, —— Powell 1963. Theoretical: ○ — ○ Raimondi 1961, numerical solution by digital computer.

throughout the wide range of parameters investigated. Experimental data by Whitley and Betts (1960), Whitley *et al.* (1962) and Powell (1963) are compared to the earlier data by Ford *et al.* (1957) in figure 5. It is seen that the later data lies below the isothermal assymptotic solution given by equation (3), and is in close agreement with Raimondi's digital computer solution.

2.2. *Whirl instabilities*

In considering the radial load capacity of self-acting gas journal bearings it is usual to assume steady-state conditions. In fact, however, there are few, if any, cases in which the gas film is steady. For example, a shaft in two journal bearings does not rotate at a steady eccentricity but orbits in the direction of rotation about this steady position due to the centrifugal effect of the out-of-balance of the shaft. This phenomenon is called synchronous whirl. The amplitude of whirl is usually small but even when the amplitude is large the motion is usually stable. It has been found that the load capacity for rotating loads is equal to the load capacity for static loads so that unbalance forces can be taken account of in the bearing design and satisfactory limits can be established for the permissible level of residual unbalance in the rotor. By dynamically balancing the rotor to within the prescribed limits it should be possible to accelerate the rotor to any desired operating speed. However in practice it is found that an upper limit of speed is set by the occurrence of a self-excited film whirl which is usually known as half-speed whirl. During half-speed whirl the shaft orbits in the bearing in the direction of rotation at half the rotational speed. The speed of transition from steady running to whirling is called the onset speed or threshold speed of half-speed whirl. If the speed is increased beyond the onset speed the amplitude of the whirl grows rapidly resulting in rubbing contact of the bearing surfaces and often in seizure. In the earlier days of gas-lubrication bearing failures occurred suddenly and unpredictably. The successful utilization of self-acting gas journal bearings became conditional upon the development of a method of predicting the onset speed of half-speed whirl.

A semi-empirical method of predicting the onset speed of half-speed whirl was described by Whitley *et al.* (1962). Whitley showed that in gas bearing machines in which the bearing sleeves were flexibly mounted four modes of whirl were possible. These are as follows:

(i) Bearing cylindrical whirl in which the bearing whirls relative to the shaft with the bearing axis and shaft axis remaining parallel.
(ii) Bearing conical whirl in which the bearing axis precesses in a conical motion relative to the shaft axis.
(iii) Shaft cylindrical whirl in which the shaft whirls relative to its bearings with its axis remaining parallel with the bearing axes.
(iv) Shaft conical whirl in which the shaft whirls with its axis precessing in a conical motion relative to the bearing axes.

Whitley found that it was shaft whirls which onset at the lowest rotational speeds in most practical machine designs. He showed that the onset speeds in the four modes were related and that if any one was measured the other three could then be calculated. He therefore proposed a method in which a freely-supported bearing sleeve was supported on a shaft which was in turn supported in two other slave bearings. The shaft was slowly accelerated until the test bearing began to whirl and the onset speed was noted. It was then possible to predict the onset speeds of shaft whirl in any machine in which the tested bearing was employed. Sensitive instrumentation allowed the whirl onset to be detected without any risk of damage to the bearing surfaces as the shaft speed was controlled to prevent the onset-speed being exceeded. Whitley's method allowed machines to be successfully produced with self-acting gas bearings but in some cases the attainment of the desired operating speed was only achieved after a lengthy period of development. There remained a need for a method of theoretically predicting the onset of half-speed whirl. Several unsuccessful attempts were made before Marsh published his theory in 1965.

The complete form of Reynolds equation for a journal bearing under non-steady conditions can be written

$$\frac{\partial}{\partial x}\left(\rho h^3 \frac{\partial P}{\partial x}\right)+\frac{\partial}{\partial y}\left(\rho h^3 \frac{\partial P}{\partial y}\right)=6\mu\omega\left\{a\frac{\partial}{\partial x}(\rho h)-\frac{2}{\omega}\rho h_0 \frac{\mathrm{d}\epsilon}{\mathrm{d}t}\cos\theta - \frac{2}{\omega}\rho h_0 \epsilon \frac{\mathrm{d}\theta_t}{\mathrm{d}t}\sin\theta+\frac{2h}{\omega}\frac{\partial \rho}{\partial t}\right\} \quad (6)$$

where $\theta_t$ is the instantaneous angle between the line of centres and a fixed radial coordinate and $\omega = U/a$ the angular velocity of the shaft.

It will be noticed that for steady-state conditions equation (6) reduces to equation (1).

Equation (6) contains a term in which density is differentiated with respect to time. Some previous attempts to develop a theory of half-speed whirl had assumed that since the instability developed from a stable condition, this term was not important. Others assumed that the Sommerfeld conditions of constant density and viscosity held for low compressibility numbers. Both assumptions lead to the wrong functional dependence of the resultant gas film force on the compressibility number. A rigorous theory for gas film lubrication in respect of dynamic performance must retain both compressibility and time-dependent effects in the Reynolds equation.

If the journal centre is considered to be rotating about the bearing centre, then a considerable simplification can be obtained by expressing Reynolds equation in a system of coordinates which rotate with the line of centres. The two coordinate systems are related by

$$\theta = \theta + \theta_t(t)$$

or

$$x = x + a\theta_t(t).$$

Equation (6) then becomes

$$\frac{\partial}{\partial x}\left(\rho h^3 \frac{\partial P}{\partial x}\right)+\frac{\partial}{\partial y}\left(\rho h^3 \frac{\partial P}{\partial y}\right)=6\mu U\left(1-\frac{2}{\omega}\frac{\mathrm{d}\theta_t}{\mathrm{d}t}\right)\frac{\partial}{\partial x}(\rho h) \quad (7)$$

which is of the same form as Reynolds equation for steady-state operation (equation (1)). However the addition term $1-(2/\omega)\,(\mathrm{d}\theta_t/\mathrm{d}t)$ is of interest since it indicates that if the journal centre follows a circular orbit around the bearing centre at one half of the rotational speed, then the pressure within the bearing clearance is constant. A steady half-speed whirling motion about the bearing centre destroys the load-carrying capacity of any journal bearing which is operating with a Newtonian lubricant. Dynamic loading at one half the rotational frequency is clearly a condition which should be avoided. Equation (7) showed that bearing failure was the probable result of half-speed whirl. It remained to discover at what condition the half-speed whirl could become self-exciting.

Marsh made only one major assumption, namely that a local linearization may be applied to the pressure field generated by small dynamic motions. The overall non-linearity of the pressure field due to static loading of the bearing was retained. A dynamic analysis of bearing behaviour was made leading to the formation of translational and conical dynamic stiffness matrices which were calculated directly from Reynolds equation. Stability criteria were then derived for cylindrical and conical modes of whirl in both one and two bearing machines. The boundary of stability or onset of whirl was shown to occur when a condition was reached for which there was no overall damping in the gas film.

For accurate stability boundaries of a bearing system it is necessary to employ a digital computer in which case an accuracy of better than 4 per cent is achieved. However, Marsh found it possible to predict whirl onset speeds to a lower but still practical accuracy by means of simple stability charts. His original work dealt only with rigidly-mounted or freely-supported bearing systems but he has subsequently developed the method to cover the case of undamped flexibly mounted bearings. Marsh has also applied his method to the prediction of whirl onset speeds in oval journal bearings.

The introduction of a small degree of ovality into the bore of the bearing sleeve is one method by which the whirl onset speed can be raised. Another method is to vent the low pressure side of the bearing to atmosphere. This method can be applied to machines with horizontal shafts in which the shaft weight is the principal bearing load and is constant in direction. Vented bearings were employed on the early gas circulators developed by the UKAEA. The effect of the vent is to raise the pressure on the low pressure side of the bearing and to cause the shaft to adopt a position of greater eccentricity in the bearing. The greater the steady-state eccentricity the higher is the whirl onset speed. This stabilizing effect can also be achieved by increasing the static radial load on the bearing as shown in figure 6 which compares Marsh's theoretical predictions with experimentally measured whirl onset speeds by Whitley (1962).

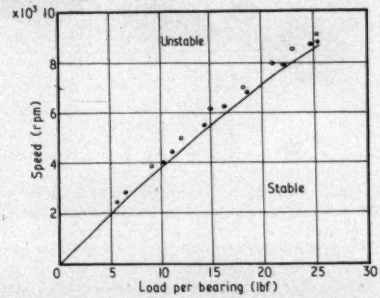



Figure 6. Cylindrical whirl onset speed as a function of load (after Marsh 1965). ○, bearing whirl mass = 25·5 lb; ●, rotor whirl, two bearings, mass = 51 lb (Whitley *et al.* 1962). —, Theory.

An extremely effective method of stabilizing self-acting gas journal bearings has been described by Kerr (1962) in which the bearing bushes are flexibly mounted in rubber 'O' rings. The rubber provides damping which permits bearings to operate without whirl up to very high speeds. Kerr has demonstrated a high frequency electric motor with a 2-in diameter shaft running in two self-acting gas journal bearings at 60 000 rpm. It is claimed that, since rubber mounted gas bearings can be stabilized with larger radial clearances than would otherwise be possible, they can be made more cheaply. It is possible that in the future they could find large-scale applications in the textile industry.

Another bearing type which has been found to be stable to very high speeds has been widely applied by Société Rateau of France. Parts of a Rateau machine are shown in figure 7. The triple pad journal bearing consists of a cylindrical bearing sleeve which has been divided into three separate pads each spanning 120° of arc and independently pivoted on ball joints at their centres of pressure. One pad is pre-loaded by means of a spring. Mech (1969) reports that machines with triple pad bearings have been operated up to a speed of 40 000 rpm with no instability. Electrically-driven gas compressors have been built with motors of up to 250 kW and with rotors weighing 150 kg. Such machines employ bearings of 16 cm or 20 cm diameter. Other machines with triple pad bearings have accumulated more than 25 000 hours of operation without any maintenance.

A third type of self-acting gas bearing which is apparently free of whirl instability is the flexible foil bearing. This bearing consists of a thin metal foil which is wrapped around the shaft and supported at the 120° positions. Foil bearings evolved from the development of high-speed tape recorders described by Gross (1967) where the magnetic tape was

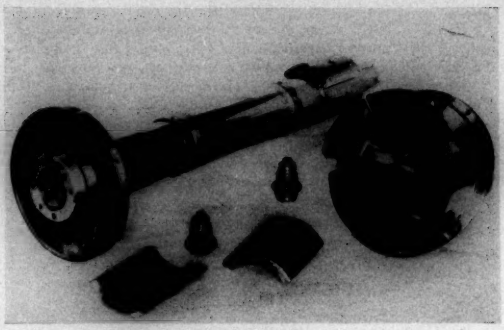


Figure 7. Tilting-pad journal and thrust bearings employed in a helium gas circulator (by courtesy of Société Rateau).

supported on a thin film of air as it passed over the recording head. It was noticed that the flexibility of the tape provided stability and offered a means of controlling self-excited instability. Foil bearings have been developed by the Ampex Corporation for use in tape recorders and other electronic equipments and by the Garrett Corporation for use in high-speed turbo compressors used in aircraft air-conditioning systems. Licht (1969) reports operating a rotor weighing 21 pounds at speeds of up to 50 000 rpm in two air-lubricated foil bearings. He observed no fractional speed whirl even with the rotor axis vertical and no radial load on the bearing. It was also found that foil bearings have a high tolerance to unbalance and can withstand temperature gradients of 60 °C across the foil width.

All of the means of delaying the onset of self-excited whirl which have so far been discussed depend upon either introducing a degree of asymmetry into the geometry or the loading of the bearing or introducing a degree of flexibility into the bearing or the bearing support. Neither of these means is acceptable in spin axis bearings for gyros (figure 8) where rigidity and iso-elasticity are essential both in the gas film and the support structure. It has become customary to design self-acting bearings for gyros with very small clearances of the range $20\text{–}100 \times 10^{-6}$ in ($5\text{–}25 \times 10^{-4}$ mm). At these small clearances the radial stiffness is high and the onset speed of self-excited whirl is raised. However, it is still often necessary to employ some other means of raising the whirl onset speed and one method which has been developed is to machine very shallow helical grooves into one of the bearing surfaces extending inwards from both ends of the bearing for between one quarter and one third of the bearing length. Gas is pumped inwards along the grooves and the pressure distribution is modified. By careful design helically grooved bearings can be made to operate at zero attitude angle which means that the deflection produced in the bearing is in the line of the applied load. This is clearly a condition conducive to stability since a change in load no longer produces an angular displacement of the rotor axis.

### 2.3. *Slip flow and turbulent flow*

In the theory of the self-acting gas bearing it is usual to assume laminar flow in the gas film with no slip at the bearing surfaces. However, as the range of application of self-acting gas bearings continues to increase a few circumstances have arisen where these assumptions need modification.

If the bearing clearance is of the same order of magnitude as the mean free path of the gas molecules then slip can occur between the gas and the bearing surfaces. In practice this is most likely to occur in bearings operating at sub-atmospheric pressures but it might

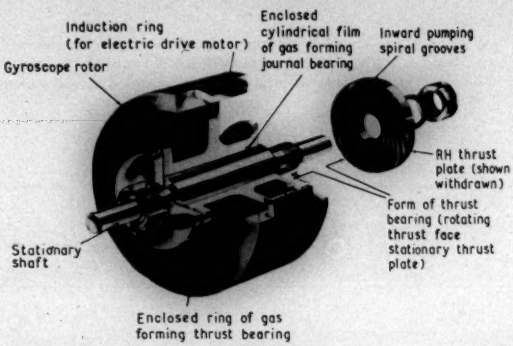



Figure 8. Gyroscope with self-acting gas bearings (by permission of the Ministry of Defence, Navy Department).

also arise in gyro bearings where the trend is to ever smaller clearances. This problem has been studied by Burgdorfer (1959) and by Wood (1968). Wood has shown that for lightly-loaded bearings ($\epsilon < 0{\cdot}5$) the load capacity of journal bearings can be calculated from the laminar flow theory by the use of a modified form of compressibility number $\Lambda'$ where

$$\Lambda' = \frac{\Lambda}{1+6m}$$

where $m = \lambda_a/h_0$ is Knudsen number and $\lambda_a$ is the mean free path at atmospheric temperature and pressure.

The modified compressibility number can be used for both steady state and dynamic performance estimations.

Wood (1968) has also considered the effect of turbulence on self-acting gas bearing performance. Turbulence may occur in large clearance bearings operating at high rotational speeds. Wood showed again that laminar flow theory with a modified form of compressibility number $\Lambda''$ could be used together with a modified length-to-diameter ratio. He showed that

$$\Lambda'' = \frac{\Lambda}{12G_x}$$

and the length-to-diameter ratio is multiplied by $(G_x/G_z)^{1/2}$ where $G_x$ and $G_z$ are functions of Reynolds number based on local clearance, local velocity and gas properties and have been tabulated by Ng and Pan (1965). At low Reynolds numbers ($R_e < 1000$) both $G_x$ and $G_z$ tend to the value $\frac{1}{12}$ and in general $G_x < G_z < \frac{1}{12}$.

### 3. Self-acting thrust bearings

Most types of machines require in addition to journal bearings some means of axially locating the rotor and withstanding thrust loading. Self-acting gas thrust bearings can take a variety of forms but consist essentially of two flat discs, one stationary and the other connected to the rotor. If both surfaces are flat and plain the rotation produces no pressure rise in the gas film and no bearing capacity is realized. It is therefore necessary to introduce some means by which the clearance between the surfaces is varied so that some pumping action occurs. The various types of gas thrust bearing differ according to the pumping method employed.

3.1. *Spirally grooved thrust bearings*

The first type of self-acting gas thrust bearing to receive serious consideration was the spirally grooved thrust bearing or viscosity plate as it is often called. In its simplest form this bearing is shown in figure 9. Both surfaces are made optically flat and one surface

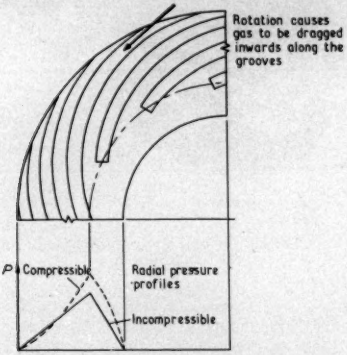



Figure 9. The spirally-grooved thrust bearing.

incorporates a series of shallow grooves extending inwards across rather more than half the radial width of the bearing annulus. The direction of rotation is such that gas is dragged inwards along the grooves with a resulting pressure rise. The gas then leaks away across the plain land. The gap between the plates, which is usually of the order $10^{-3}$–$10^{-4}$ in ($2{\cdot}5\times10^{-2}$–$2{\cdot}5\times10^{-3}$ mm), is self adjusting to the point where the integrated pressure balances the applied load. The viscosity grooves are of the same order of depth as the bearing clearance. They can be inward pumping, as shown in figure 9, or outward pumping from the inner edge of the annulus, or both inward and outward pumping grooves can be used when the bearing is called a 'herringbone' bearing.

The theory of the spirally-grooved thrust bearing was first developed by Whipple (1958) and was extensively described by Whitley and Williams (1959). In thrust bearing analysis it has always been considered valid to assume isothermal conditions and for many practical purposes compressibility can be ignored. A compressibility number can be defined which corresponds to that used in journal bearings as

$$\Lambda=\frac{6\mu U(r_0-r_1)}{h^2P_a}$$

where $r_0$ is the outer radius of bearing, $r_1$ the inner radius of bearing and $U$ the velocity at outer radius of bearing.

This form of compressibility number is inconvenient because it depends upon the gas film thickness $h$ which varies with the applied load. Whitley (1967) uses a compressibility number $\lambda$ based upon the groove depth $d$ where

$$\lambda=\frac{6\mu U(r_0-r_1)}{d^2P_a}. \tag{8}$$

Whitley following Whipple showed that the pressure ratio across the grooves developed

by the pumping action was a function of the compressibility number $\lambda$ expressed by equation (9)

$$(\tau-1)+\frac{(\tau^2-1)}{C\lambda}\ln\left\{\frac{\tau-(\tau^2-1)/C\lambda}{1-(\tau^2-1)/C\lambda}\right\}=D\lambda, \tag{9}$$

where $\tau=P_i/P_a$ the pressure ratio across the grooves and $C$ and $D$ are functions of the groove dimensions.

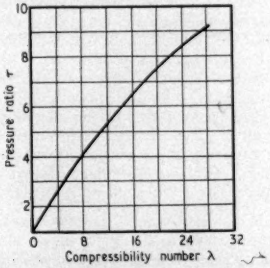



Figure 10. Effect of compressibility on pressure ratio for narrow-width spirally-grooved thrust bearing.

The solution of equation (9) is plotted in figure 10. It can be seen that the pressure ratio initially increases linearly with compressibility number but then less rapidly. It does not, however, reach an asymptotic finite level as for journal bearings but continues to increase indenfinitely with the pressure ratio increasing with the square root of $\lambda$ at high compressibility numbers. For designing to an accuracy of better than 10% incompressible theory can be used up to a pressure ratio of 6 and a compressibility number of 12.

Theories for spirally-grooved thrust bearings have also been developed by Malanoski and Pan (1964) and by Muijderman (1966). The theories are in close agreement with Whipple's. Table 1 compares the optimum groove parameters for maximum load capacity as given by Whipple and Muijderman.

The major development problems which have arisen with spirally-grooved thrust bearings have both resulted from the small clearances at which they operate. Except in very small machines such as gyros, it has proved almost impossible to rigidly mount both bearing

**Table 1. Comparison of groove optima for maximum load capacity in self-acting spirally-grooved thrust bearings.**

| Groove parameter | Outward pumping $r_0/r_1=2$ | | Inward pumping $r_0/r_1=2$ | |
|---|---|---|---|---|
| | Whipple | Muijderman | Whipple | Muijderman |
| Length $\alpha$ | 0·75 | 0·74 | 0·72 | 0·74 |
| Width $x$ | 2·29 | 2·33 | 1·68 | 1·59 |
| Depth $y$ | 4·22 | 4·45 | 3·93 | 4·04 |
| Angle $\theta$ | 70·5 | 70·7 | 72·0 | 72·5 |

where $\alpha=(r_0-r_b)/(r_0-r_i)$, $r_0$ is the outer radius, $r_b$ the radius between plain and grooved annuli, $r_i$ the inner radius of bearing; $x=b_g/b_l$, $b_g$ is the width of groove and $b_l$ the width of land between grooves and $y=(h+d)/h$, $h$ is the design clearance and $d$ the groove depth. $\theta$ is the angle between grooves and radial coordinate.

surfaces to the required degree of precision. For this reason it has become customary to mount the static plate on a gymbal mounting to permit it to self-align. However, the freedom of movement permits angular vibrations of the static plate to be driven by any swash in the movement of the rotating plate. This phenomenon has been reported by Whitley (1964) who defined critical speeds and tolerable limits of vibration.

The second major problem with spirally-grooved thrust bearings has arisen from thermal distortion of the plates caused by heat generated from the shearing of the gas film. About a quarter of the heat generated flows in an axial direction through the plates and causes a temperature gradient which distorts the plates convex towards the gas film. One method of overcoming this problem is to manufacture the plates slightly concave so that they become flat at the operating condition. Another technique is to make the plates of a highly-conducting material such as molybdenum. Alternatively the backs of the plates can be insulated so that the majority of the heat flows radially.

Spirally-grooved thrust bearings have been applied far more widely than any other type of self-acting gas thrust bearing. They are used extensively in gyros and in the majority of gas circulators developed for nuclear applications. Their popularity is largely due to their relatively high load capacity. Pantall (1964) has compared the load capacity of three types of self-acting thrust bearing all with the following dimensions and operating conditions:

Outside diameter 4 in — Inside diameter 2 in
Clearance 0·0002 in — Speed 10 000 rpm
Lubricant—air at atmospheric conditions.

He calculated the following load capacities:

| | |
|---|---|
| Spirally-grooved bearing | 126·5 lb |
| 360° stepped sector bearing | 66·5 lb |
| Tilting pad bearing | 29·1 lb |

### 3.2. *Other types of self-acting thrust bearings*

Figure 11 shows the stepped sector thrust bearing and the pocket thrust bearing. These two types have received some attention but comparatively little application. Design data for the stepped sector thrust bearing has been derived by Ausman (1961). This bearing has the advantage of being relatively easy to manufacture but its load capacity is reduced by the ease by which the gas can flow radially to atmosphere thus reducing the pressure build-up over the step. The pocketed bearing is a refined version of the stepped sector thrust bearing in which the sides of the step are bounded by narrow lands to prevent radial leakage. This bearing is more difficult to manufacture. Experimental and application details have been reported by Drescher (1959) who used pocketed thrust bearings on a 60 kW motor alternator set.

A more serious rival to the spirally-grooved thrust bearing is the tilting-pad thrust bearing. In this bearing the annulus is divided into a number of segments each of which is

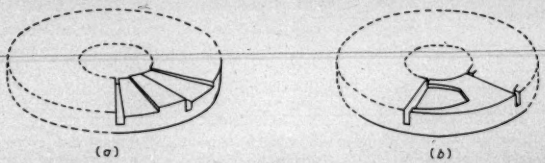



Figure 11. Types of gas thrust bearing. (*a*) Stepped sector thrust bearing. (*b*) Pocketed thrust bearing.

individually supported on a ball joint or pivot. Six pads are commonly used. The bearing surfaces are made slightly convex to prevent film collapse because if a flat pad becomes parallel with the rotating face there is no change in the gas film thickness to cause pressure generation and hence no load capacity. A theoretical study and some experiments were reported by Gross *et al.* (1959) and Gross (1962). Gross showed that the load capacity is a function of the compressibility number based on the minimum clearance and the relationship is similar to that for cylindrical journal bearings. Figure 12 shows the curves of load capacity with compressibility number derived by Gross for convex cylindrical pads. He recommended an optimum crown height of 0·0003 in (0·0075 mm) and an optimum pivot position which is 0·6 of the length from the leading edge of the pad.

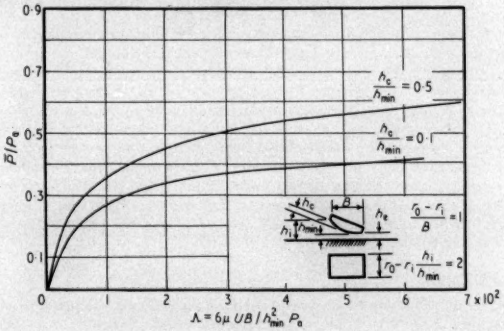



Figure 12. Tilting-pad thrust bearing: load capacity for convex cylindrical pad.

Pivoted-pad thrust bearings have been mostly employed in combination with pivoted-pad journal bearings. Mech (1969) has described their use in turbo-machinery constructed by Société Rateau. Compared to the other types of self-acting thrust bearing they are easier to manufacture since they do not incorporate any intricate grooves or pockets. The pads can be faced with self-lubricating material containing, for example, PTFE, to withstand rubbing contact at start-up before hydrodynamic action becomes established, and the self-aligning ability of the pads can accommodate some distortion of the rotating plate that might arise in operation. These factors confer on the tilting-pad bearing advantages of economy and ruggedness which can more than offset their disadvantages in terms of load capacity and stiffness in some applications. Parts of an early Rateau machine with tilting-pad thrust and journal bearings are shown in figure 7.

## 4. Externally pressurized gas journal bearings

The method of operation of an externally pressurized journal bearing is explained with reference to the typical jet-fed journal bearing of figure 13. The bearing consists of a cylindrical bush into which are drilled two rows of gas feed holes spaced evenly around the bearing circumference. Compressed gas from an external source is supplied to the reservoir surrounding the bearing. From the reservoir the gas flows through the feed holes into the clearance between the shaft and the bush and then axially to the ends of the bearing where it exhausts to atmosphere. The pressure in the reservoir is the gas supply pressure $P_0$. The pressure falls as the gas flows through the feed holes and enters the bearing clearance at a pressure $P_d$, exhausting at the end of the bearing at a pressure $P_a$.

With no load applied to the shaft (and neglecting its weight), the shaft adopts a concentric

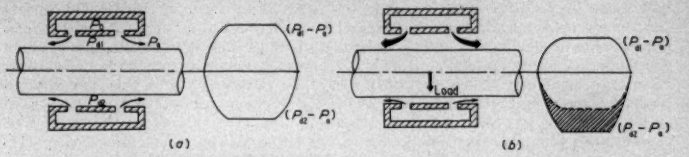



Figure 13. The externally-pressurized journal bearing. (*a*) No load on shaft. Nureo press difference. (*b*) Load applied to shaft. Pressure difference supporting load.

position in the bush. In this case there is no variation of pressure circumferentially around the bearing and $P_{d1}=P_{d2}$. The pressure forces on the shaft balance,

When a load is applied to the shaft in the vertically downward direction, the shaft deflects in that direction so that the clearance at the top of the bearing is increased, and that at the bottom is reduced. At the top of the bearing the resistance to the flow of gas escaping to atmosphere is thus lessened so that more gas flows in through the top feed holes. This flow increases the pressure drop through the feed holes and the pressure $P_{d1}$ falls. At the bottom of the bearing the resistance to the flow of gas escaping to atmosphere has been increased so less gas flows through the bottom feed holes. The pressure drop through these feed holes has been reduced and the pressure $P_{d2}$ rises. A difference of pressure now exists across the shaft to balance the applied load. For all loads within the capacity of the bearing there is an equilibrium position for the shaft.

### 4.1. *Static load capacity*

In the study of self-acting bearings one is concerned with the viscous flow of gas in a narrow slot in which the flow is induced by the motion of one of the surfaces. In externally pressurized bearings one is primarily concerned with flow which is induced by pressure. Figure 14 illustrates the flow of gas in a narrow slot in which the pressure falls from $P_1$ to $P_2$

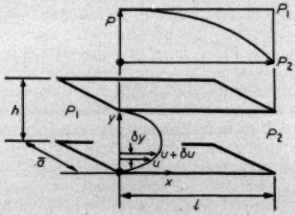



Figure 14. Pressure-induced flow between parallel plates.

along the length of the slot $l$ in the $x$ direction. Applying the same assumptions as before the Navier–Stokes equations reduce to

$$\frac{\partial^2 u}{\partial y^2}=\frac{1}{\mu}\frac{\partial P}{\partial x}$$

for flow only in the $x$ direction. Integrating twice and applying the boundary condition $u=0$ at $y=0$ and $y=h$ gives the gas velocity $u$ at any plane $y$ in the film as

$$u=\frac{1}{2\mu}\frac{\mathrm{d}P}{\mathrm{d}x}y(y-h).$$

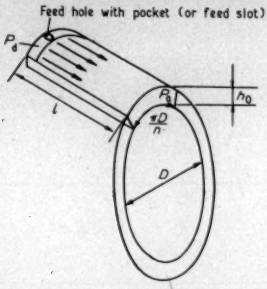



Figure 15. Equivalent slot in a journal bearing.

Integrating across the film gives the mass flow rate of gas $m$ as

$$m = -\frac{\bar{a}\rho h^3}{12\mu}\frac{\mathrm{d}P}{\mathrm{d}x}$$

and separating the variables and integrating finally yields

$$P_1^2 - P_2^2 = \frac{24\mu mRTl}{\bar{a}h^3} \tag{10}$$

for isothermal conditions in the gas film. Equation (10) gives the relationship between pressure drop, mass flow rate, dimensions and gas properties for a rectangular slot. In journal bearing analysis it is usual following Shires (1949) and Robinson and Sterry (1958) to divide the bearing into a number of 'equivalent slots' each supplied with gas from a feed hole or feed slot. This arrangement is shown in figure 15. The slot width is $\bar{a} = \pi D/n$ where $D$ is the diameter of the bearing and $n$ is the number of feed holes in a row around the bearing. When the shaft is concentric in the bearing the clearance in all the slots is the mean radial clearance of the bearing $h_0$. Usually there are two rows of feed holes supplying slots running to the two ends of the bearing. For such a bearing the total mass flow rate of gas through the bearing is obtained from equation (10) as

$$M = \frac{(P_d^2 - P_a^2)\,\pi D h_0^3}{12\mu RTl} \tag{11}$$

where $P_d$ is the pressure immediately downstream of the feed hole and $P_a$ is the ambient pressure at the bearing exhaust. It is seen that the gas flow is proportional to the cube of the clearance and this dependence provides a strong inducement to design bearings of the smallest practicable clearance.

In order to attempt to predict the load capacity of the externally pressurized journal bearing it is necessary to consider the feed jet or feed slot in combination with the bearing slot. In considering the gas flow through a feed hole it is usual to assume adiabatic conditions because the high velocity and short jet length permit little opportunity for heat transfer. The flow through the jet is given by

$$m = C_D A \rho_0 (2RT_0)^{1/2} F(\gamma K) \tag{12}$$

where $C_D$ is the coefficient of discharge, $A$ is the cross section area of the throat of the jet, $\rho_0$ and $T_0$ are the density and temperature at the supply conditions.

$$F(\gamma K)=\left[\frac{\gamma}{\gamma-1}\left\{\left(\frac{P_d}{P_0}\right)^{2/\gamma}-\left(\frac{P_d}{P_0}\right)^{\gamma+1/\gamma}\right\}\right]^{1/2}$$

in which $K=P_d/P_0$ and $P_0$ is the gas supply pressure.

Both $F(\gamma K)$ and $C_D$ vary with $K$. However, $C_D$ is often taken to have a constant value of 0·8 which is true for $0<K<0{\cdot}7$. Similarly $F(\gamma K)$ becomes constant for choked feed hole conditions and for air $F(\gamma K)=0{\cdot}484$ for $0<K<0{\cdot}528$. Combining equations (10) and (12) for a feed hole supplying one equivalent slot and rearranging yields

$$K_g=\frac{24\mu(2RT)^{1/2}\,lC_DAF(\gamma K)}{(P_d/P_0+P_a/P_0)\,(1-P_a/P_0)\,P_0\bar{a}h^3} \tag{13}$$

where $K_g=(P_d-P_a)/(P_0-P_a)$ is called the gauge pressure ratio.

The calculation of load capacity of a journal bearing requires the solution of equation (13) for each slot around the circumference. When load is applied to the shaft it moves to an eccentric position in the bearing and the clearance at an angle $\theta$ to the load line is given as before as

$$h=h_0(1-\epsilon\cos\theta).$$

Thus for any given mean radial clearance and eccentricity ratio the mean clearance for each slot can be calculated and substituted into equation (13) to obtain the pressure at which the gas enters the bearing. Then assuming that the pressure falls parabolically to the ends of the bearing, and that the pressure is constant in the centre of the bearing between adjacent jets, the pressure forces of the slots on the shaft can be summed vectorially to yield the load on the shaft.

Equation (13) is not explicit in terms of $P_d$ or $K_g$ and hence its solution is a process of successive approximations which are lengthy and tedious. However, Shires obtained an explicit relationship for incompressible flow

$$K_g=\frac{2}{1+(1+4/G^2)^{1/2}} \tag{14}$$

where the slot factor $G$ is given by

$$G=\frac{P_a/P_0}{(1+P_a/P_0)\,(1-P_a/P_0)^{1/2}}\,\frac{24\mu(2RT)^{1/2}}{P_a}\,\frac{lC_DA.}{\bar{a}h^3}$$

He showed that the incompressible relationship between $K_g$ and $G$ closely approximated the compressible relationship. It is then only necessary to make one calculation of $G_0$ for the concentric bearing. Then for any slot at angle $\theta$ to the load line

$$\frac{G}{G_0}=\left(\frac{h_0}{h}\right)^3=\frac{1}{(1-\epsilon\cos\theta)^3}$$

and the corresponding value of $K_g$ is given from equation (14). In this way the pressure distribution in the bearing is found for any eccentricity ratio and the corresponding load capacity can be calculated. The gauge pressure ratio for a concentric journal bearing $K_{g0}$ is an important design parameter. The variation of load coefficient $C_L$ with $K_{g0}$ as calculated by the method described is shown in figure 16 for eccentricity ratios of 0·1, 0·5 and 0·9

$$C_L=\frac{\bar{P}}{P_0}$$

where $\bar{P}$ is the load per unit bearing plan area as before. Figure 16 gives the load capacity of a bearing in which the gas flows axially from the plane of the feed holes to the bearing ends. In practice however this axial flow model is only approached in a short bearing with many

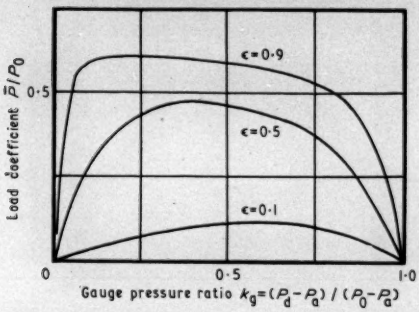



Figure 16. Theoretical load capacity of short externally pressurized journal bearings.

feed holes. In a real bearing of practical length-to-diameter ratio and with a practical number of feed holes the axial flow is disturbed by the effects of dispersion and circumferential flow shown in figure 17. Dispersion is the divergence of the gas flow from its point of entry to the gas film. Its effect on load capacity and flow has been studied by Dudgeon and Lowe (1965). They showed that the effect was greatest where few feed holes were used and where the feed holes were close to the ends of the bearing.

Circumferential flow is induced by the circumferential pressure variations which arise in an eccentric bearing. The gas tends to flow around the shaft from the high pressure region to the low pressure region partly destroying the pressure difference and reducing the radial load capacity. No simple analytical solution has been found for predicting the effects of circumferential flow. Dudgeon and Lowe have given a digital computer solution and Shires *et al.* (1963) has devised a semi-empirical expression based on experimental data by Robinson and Sterry. The loss of load capacity increases with length-to-diameter ratio.

Externally pressurized journal bearings in which the gas is supplied to the clearance through feed slots can be designed so that the gas enters through almost the entire 360 degrees of circumference. In such bearings dispersion effects are non-existent. Figure 18 shows a journal bearing with circumferential inlet slots. The analysis of slot fed bearings was

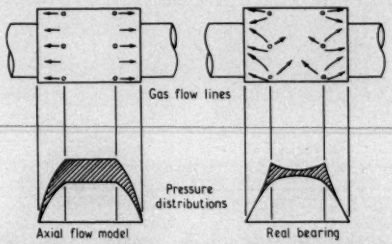



Figure 17. The effects of dispersion and non-axial flow.

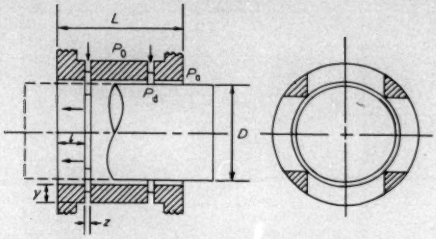



Figure 18. Journal bearing with circumferential inlet slots.

shown by Shires and Dee (1967) to be much simpler than for jet fed bearings. In a bearing with $n$ feed slots of radial length $y$ and thickness $z$ equation (10) can be written

$$P_0^2 - P_d^2 = \frac{24\mu RTmy}{z^3}\frac{n}{\pi D}$$

and the flow in the equivalent slot is given by

$$P_d^2 - P_a^2 = \frac{24\mu RTml}{h^3}\frac{n}{\pi D}.$$

Eliminating the mass flow $m$, and at the same time $\mu$, $R$, $T$, $D$ and $n$ gives

$$\frac{P_d^2 - P_a^2}{P_0^2 - P_a^2} = \frac{1}{1+\left(\frac{y}{l}\right)\left(\frac{h}{z}\right)^3} = \frac{1}{1+\alpha}.$$

Putting $\beta = 2P_a/(P_0 - P_a)$ and rearranging yields

$$K_g^2 + \beta K_g - \frac{1+\beta}{1+\alpha} = 0$$

which has a real solution

$$K_g = \left\{\left(\frac{\beta}{2}\right)^2 + \frac{1+\beta}{1+\alpha}\right\}^{1/2} - \frac{\beta}{2}. \tag{15}$$

It may be noted that for a single ring of feed slots around the centre of the bearing

$$\alpha = 2\left(\frac{y}{l}\right)\left(\frac{h}{z}\right)^3$$

and for two rings of slots

$$\alpha = \left(\frac{y}{l}\right)\left(\frac{h}{z}\right)^3$$

because in the former case one feed slot supplies two bearing slots. Equation (15) can be used to determine the pressure $P_d$ for each slot and load capacity can be derived as described for a bearing with feed holes. Figure 19 illustrates the load capacity of short slot fed bearings; that is neglecting the effect of circumferential flow.

One significant feature of all bearings with inlet slots is that the load capacity and stiffness are functions of the dimensions and fluid pressures only. Unlike bearings with metering orifices, gas properties and temperature do not influence the performance. Furthermore

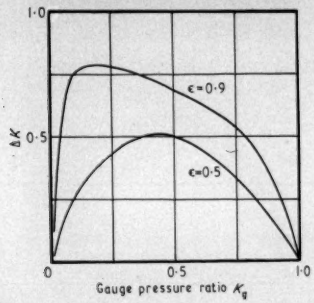



Figure 19. Load capacity of short journal bearings with circumferential inlet slots (after Shires and Dee 1967). Calculated for 6 entries, $P_0 = 3P_a$, load coefficient $C_L = (1 - 1/L)\,\Delta K$.

the choice of optimum dimensions, is not significantly affected by the level either of the supply pressure or of the ambient pressure over the normal working ranges. In fact a bearing designed for pressurizing with a gas will also work at the same gauge pressure ratio with a liquid. The maximum radical stiffness is achieved when $\alpha = 8$.

Shires and Dee (1967) have also reported an analysis of journal bearings with axial feeding slots. The gas flow is subject to dispersion effects but it has been suggested that by continuing the slots to close to the ends of the bearing the effect of circumferential flow can be minimized. This has yet to be demonstrated experimentally, and, in view of the manufacturing complexity, it is difficult to foresee applications in which the use of axial slot feeding would be advantageous.

### 4.2. *Steady-state dynamic performance*

In considering the static load capacity of externally pressurized gas journal bearings all flow is assumed to be pressure induced and surface velocity effects are neglected. This condition occurs in practice when the shaft is stationary. The load capacity and stiffness of the stationary bearing are called the 'aerostatic' load capacity and stiffness and they can be simply measured by applying load to the stationary shaft and measuring the resulting radial deflection. Rotation of the shaft modifies the gas flow in the bearing clearance and the resulting pressure changes affect the load capacity and stiffness of the bearing. When the velocity effects become significant the bearing is said to be operating in a hybrid mode and bearings designed to operate in this mode are called hybrid bearings.

An experimental investigation of the steady-state operation of a hybrid gas journal bearing was reported by Powell (1964). It was shown that as the shaft accelerated under conditions of constant radial load the shaft centre moved progressively to positions of lower eccentricity and out of the line of the applied load. The bearing exhibited the characteristics found in self-acting bearings of increasing load capacity and the adoption of an attitude angle. The changes of eccentricity ratio and attitude angle with speed are shown in figure 20. The steady-state load capacity and stiffness increased with speed to an extent that doubled or even trebled the aerostatic values. Figure 21 shows data obtained from load-deflection tests carried out at zero speed and three finite speeds. It is seen that for the test bearing the radial load capacity at an eccentricity ratio of 0·6 increased from 26 lb (12 kg) at zero speed to 84 lb (38 kg) at 7 500 rpm.

Powell (1964) has shown that the effects of shaft rotation in hybrid journal bearings can

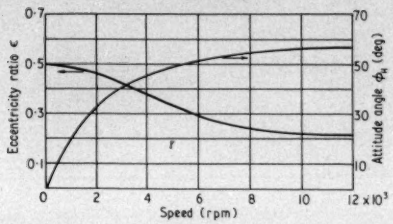



Figure 20. Hybrid gas journal bearing—variation of eccentricity ratio and attitude angle with speed.

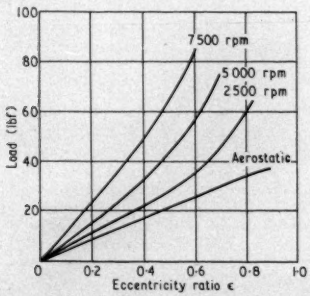



Figure 21. Hybrid gas journal bearing—variation of load capacity with eccentricity ratio and speed.

be related to a compressibility number $\Lambda_m$ where

$$\Lambda_m = \frac{6\mu a U}{P_m c^2}$$

and $P_m$ is the mean pressure in the gas film and for most practical bearings $P_m = (P_0 + P_a)/2$. Values of $\Lambda_m$ are usually low enough to assume that the velocity effects are occurring in the near incompressible region. It was suggested that the load capacity of hybrid journal bearings could be predicted from self-acting and externally-pressurized theory by linear superposition provided that an empirical factor was obtained to reduce the self-acting effect.

The interaction of the two effects can be visualized with the aid of the displacement and force diagrams shown in figure 22. The shaft rotating with angular velocity $\omega$ about its centre O′ is deflected by a load $W$ a distance $OO' = \epsilon h_0$ from the centre of the bearing O. The deflection occurs at some angle $\phi_H$ (hybrid attitude angle) to the direction of the applied load. The resultant pressure force acting on the shaft is equal and opposite to the applied load. This resultant force is made up of two components, the aerostatic pressure field and the pressure changes produced by shaft rotation. The aerostatic pressures produce a force $W_S$ which opposes the shaft deflection. The effect of rotation produces a force $W_d$ at some angle $\phi$ to the deflection. Solving the force triangle yields:

$$W = W_S \cos \phi_H + W_d \cos (\phi - \phi_H) \tag{16}$$

and

$$\cot \phi_H = \frac{W_S}{W_d \sin \phi} + \cot \phi. \tag{17}$$

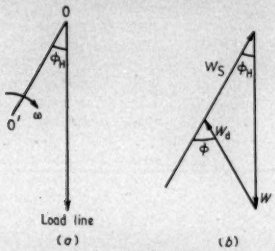



Figure 22. Quasi-static equilibrium in a hybrid journal bearing. (*a*) Displacement diagram. (*b*) Force diagram.

$W_d$ can be derived from self-acting journal bearing design curves of $\bar{P}/P_a$ against $\Lambda$. Substituting $P_m$ for $P_a$ and $\Lambda_m$ for $\Lambda$ one has

$$W_d = E\frac{\bar{P}}{P_m} LDP_m$$

where $\bar{P}/P_m = \bar{P}/P_a$ for $\Lambda = \Lambda_m$ and $E$ is an empirical factor which Powell (1970) gives as 0·7 for bearings of length-to-diameter ratio of 2. Similarly, self-acting bearing design curves give graphs of $\phi$ against $\Lambda$ and the appropriate value of $\phi$ is obtained by putting $\Lambda = \Lambda_m$. The method of linear superposition is useful in visualizing the operation of hybrid bearings and predicting the trend of performance with variation of such design parameters as speed, clearance and gas supply pressure. However there is little justification for the separation of the two effects and the reliable range of parameter variation in the design of bearings has not been fully established. Lund and Vohr (1967) refer to digital computer predictions of the steady-state performance of hybrid bearings and show good agreement with experimental data by Powell. These computer aided design methods have been developed by Mechanical Technology Incorporated of Latham, New York.

### 4.3. *Whirl instabilities*

The whirl instabilities which occur in hybrid gas journal bearings are similar to those described for self-acting bearings. However, because the externally pressurized bearing possesses aerostatic stiffness which is independent of shaft rotation and remains effective when self-acting effects break down as, for example, during half-speed whirl, hybrid bearings possess natural frequencies determined by the aerostatic stiffness and the mass and moments of inertia of the rotor. For a symmetrical system of a rotor supported in two identical journal bearings the lowest natural frequency $\omega_1$ in the cylindrical mode is given by

$$\omega_1^2 = \frac{2K}{m_R} \tag{18}$$

where $K$ is the aerostatic radial stiffness of one journal bearing and $m_R$ is the mass of the rotor.

The lowest natural frequency $\omega_2$ in the conical mode is given by

$$\omega_2^2 = \frac{2KJ^2}{I - I_0} \tag{19}$$

where $2J$ is the distance between bearing centres, $I$ is the transverse moment of inertia of the rotor, and $I_0$ is the polar moment of inertia of the rotor.

Most of the early investigators of externally pressurized bearings tested bearings of relatively large clearance compared to contemporary self-acting bearings. These large clearance bearings (roughly $h_0/a > 0{\cdot}002$) exhibited dynamic characteristics which appeared to be quite distinct from those found in self-acting bearings. Fischer *et al.* (1959) reported that synchronous whirls were excited by residual unbalance in the rotor and that these passed through resonances of large amplitude at one or both of the natural frequencies defined by equations (18) and (19). Above these speeds the rotor appeared to have inverted to rotation about its mass axis and the unbalance forces diminished to negligible proportions. Further acceleration revealed no change in this condition until suddenly the onset of unstable self-excited whirl occurred. For lightly-loaded bearings the onset speed corresponded to twice the lower natural frequency and the whirl frequency was half the rotational frequency. Thus half-speed whirl was found to occur in which the whirl frequency corresponded to the lower natural frequency of the system. Powell and Tully (1962) found that loading the bearings, for example by adding load to the shaft through a centrally-located slave bearing, raised the whirl onset speed. However, the onset speed always remained an integer multiple of the lower natural frequency and the whirl frequency was always equal to the lower natural frequency. Thus they encountered one-third speed whirl, one-quarter speed whirl and so on up to one-seventh speed whirl.

Tully (1966) showed experimentally for externally pressurized bearings, as Marsh had shown for self-acting bearings, that self-excited whirl onsets when the damping in the gas film becomes zero. Tully measured the natural frequency and damping of a two-bearing system by observing the decaying oscillations of the rotor following a shock loading. He repeated the measurements at several speeds up to the whirl onset speed. The results are shown in figure 23 for four gas supply pressures. It is seen that the natural frequency

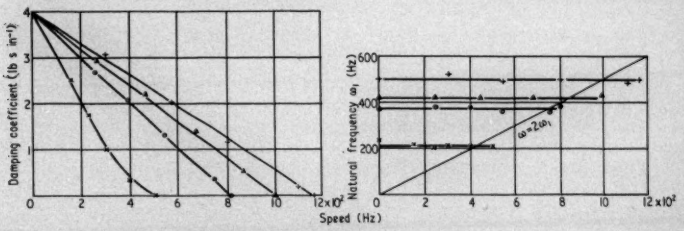



Figure 23. Variation of damping and natural frequency with speed for large clearance bearing (after Tully 1966). $D = 0{\cdot}75$ in, $L = 1{\cdot}50$ in, $h_0 = 0{\cdot}0011$ in. ×, $P_0 - P_a = 10$ lbf in$^{-2}$; △, $P_0 - P_a = 50$ lbf in$^{-2}$; ○ $P_0 - P_a = 30$ lbf in$^{-2}$; +, $P_0 - P_a = 80$ lbf in$^{-2}$.

remains constant but the damping falls linearly with speed. Whirl onsets when the damping becomes zero at a speed close to twice the natural frequency. The damping present when the shaft is not rotating but is forced to vibrate translationally by, for example, a shock load, is derived from squeeze film action. For a segment of an infinitely long bearing the dimensionless squeeze film damping is given by

$$W_{sf} = \frac{6\pi}{(1-\epsilon^2)^{3/2}} \frac{r}{h_0} \left(\frac{\Lambda}{\omega}\right) \omega^* \tag{20}$$

where $r$ is the half amplitude of the vibration, $\Lambda = (6\mu\omega/P_a)(a/h_0)^2$ and $\omega^*$ is the vibration frequency.

When the shaft is rotating, self-acting forces arise in the bearing. If the clearance ratio is large the compressibility number is low and the operation is in the near incompressible region. During vibrations of the shaft the self-acting force acts at 90° to the displacement

and constitutes negative damping which lessens the overall damping of the bearing. The dimensionless self-acting damping force for a segment of an infinitely long bearing is given by

$$\overline{W}_{\mathrm{d}} = \frac{6\pi}{(2+\epsilon^2)(1-\epsilon^2)^{1/2}} \frac{r}{h_0} \left(\frac{\Lambda}{\omega}\right) \omega \tag{21}$$

and the overall damping force $\overline{W}$ is given by

$$\overline{W} = \overline{W}_{\mathrm{st}} - \overline{W}_{\mathrm{d}}.$$

In a large clearance bearing operating in the near incompressible region the radial stiffness remains effectively independent of the rotational speed. Hence the natural frequency $\omega^*$ which ensues following a shock loading also remains independent of speed. Under these conditions it can be seen that the squeeze film damping as defined by equation (20) remains constant and independent of rotational speed. The self-acting force however increases linearly with speed and hence the net damping falls linearly with speed as shown by Tully's experiments.

Tully also showed that bearings of smaller clearance possess considerably higher damping and behave dynamically rather more like self-acting bearings. Some experimental data is shown in figure 24. The natural frequency falls with increasing speed until it becomes

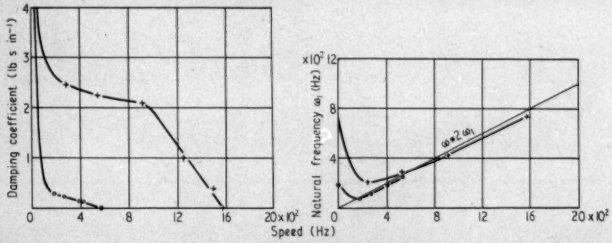



Figure 24. Variation of damping and natural frequency with speed for low clearance bearing (after Tully 1966). $D=0{\cdot}75$ in, $L=1{\cdot}50$ in, $h_0=0{\cdot}0005$ in. ○, $P_0-P_a=50$ lbf in$^{-2}$. +, $P_0-P_a=80$ lbf in$^{-2}$.

equal to half the rotational speed. It then increases linearly maintaining approximately the half-speed relationship until the self-excited whirl onset speed is reached. However the onset speed is still close to twice the natural frequency, based on the aerostatic stiffness that is measured at zero speed. The damping falls rapidly at low speeds, reaches a plateau and then falls linearly as the whirl onset speed is approached. Bearings of small clearance do not exhibit synchronous whirl resonances determined by the aerostatic stiffness. In the low speed region where the natural frequency is falling a point is passed through where the natural frequency coincides with the rotational frequency. At this point a synchronous resonance is sometimes observed but the high damping in this region often limits the whirl amplitude to negligible proportions.

Although the understanding of the dynamic behaviour of hybrid bearings remains incomplete sufficient is known for many design purposes. The onset speed of self-excited whirl can be predicted and limits can be established for the residual unbalance of the rotor. Where a particular degree of damping is required, as for example in some machine tool applications, this can be determined at the design stage and adjusted during the development period. Tully showed that where bearings incorporate pockets in the gas film around the feed holes, the larger the pocket volume the lower is the initial damping. Thus by variation of the pocket depth a desired damping coefficient can be achieved.

In some applications of hybrid bearings it is not possible to design the bearings to ensure that the self-excited whirl onset speed is above the operating speed range. In such cases it is necessary to employ one of several methods which have been developed to suppress whirl. One of the earliest was described by Sixsmith (1959). Sixsmith used a series of cavities arranged radially around the centre of the bearing sleeve and each vented to atmosphere through a hole, the diameter of which was adjusted to tune out the whirl frequency. This method was employed successfully on small expansion turbines operating at speeds of up to 250 000 rpm. These units were used in helium liquefaction plants with the bearings operating on helium gas at temperatures down to 12 K.

Hybrid gas journal bearings can be stabilized for operation at very high speeds by flexibly mounting the bearing sleeves in rubber rings. This method was employed in one of the earliest widescale applications of air lubricated bearings shown in figure 25. The air

Figure 25. Sectioned air-bearing dental turbine showing air passages, turbine and rubber stabilized air bearings (by courtesy of the Dental Manufacturing Co. Ltd).

bearing dental turbine was introduced in 1962 and since then over 30 000 have been manufactured and sold throughout the world. These units operate on compressed air supplied at 50 lbf $in^{-2}$ (3·5 kg $cm^{-2}$) and run at a speed of 500 000 rpm. Powell (1968) has described the rubber stabilization of high-speed bearings and has drawn attention to the need to select the correct rubber in terms of hardness, chemical compatibility and heat resistance. In general, the smaller the machine the softer the rubber that must be used. Butyl and nitrile rubbers possess high damping and can be used at temperatures up to about 60 °C. At higher temperatures up to 200 °C silicone or Viton rubbers are effective.

## 5. Externally pressurized thrust bearings

Figure 26 shows the simplest form of externally pressurized gas thrust bearing in which a circular pocket is fed by a single central feed hole. The gas flows radially outwards from the edge of the pocket to the edge of the bearing. Assuming laminar flow and conditions as before the Navier–Stokes equations reduce to:

$$\frac{\partial^2 u}{\partial y^2} = \frac{1}{\mu}\frac{\partial P}{\partial r}.$$

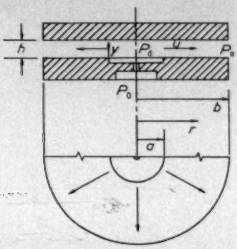



Figure 26. The simple aerostatic thrust bearing with central feed hole and pocket.

Integrating twice and substituting for the mass flow $m$ gives

$$\frac{dP}{dr} = -\frac{6\mu m}{\pi r \rho h^3}$$

and assuming isothermal conditions and integrating again yields the pressure $P$ at any radius $r$ in the gas film.

$$P_d^2 - P^2 = \frac{12\mu m RT}{\pi h^3} \ln\left(\frac{r}{a}\right)$$

where $P_d$ is the pressure in the pocket.

This simple thrust bearing can be used occasionally in practice but a more common type of bearing is shown in figure 27. The annular thrust bearing has a central hole to allow a

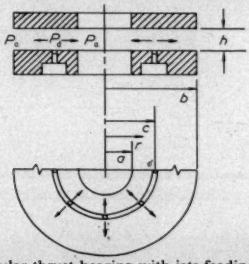



Figure 27. The annular thrust bearing with jets feeding an annular groove.

shaft to pass through. It is supplied with gas by a ring of jets feeding an annular groove or alternatively by an annular slot. In such a bearing the pressure outside of the feeding ring is given by

$$P_d^2 - P^2 = \frac{12\mu m_1 RT}{\pi h^3} \ln\left(\frac{r}{c}\right)$$

and the pressure inside the feeding ring is given by

$$P_d^2 - P^2 = \frac{12\mu m_2 RT}{\pi h^3} \ln\left(\frac{c}{r}\right)$$

where $c$ is the radius of the feeding ring.

It will be noticed that for equal outwards and inwards radial flow ($m_1 = m_2$), $c^2 = ab$ where $a$ is the inside radius, and $b$ is the outside radius of the bearing.

If the pressure is integrated over the bearing area the load capacity of the bearing can be found. Unfortunately the integration of the compressible flow equation yields expressions for load capacity which are difficult to evaluate. However, the effect of compressibility on the pressure distribution in most thrust bearings is small and the incompressible pressure distributions can be used to give the load capacity $W$ of the simple thrust bearing as:

$$W = K_g(P_0 - P_a)\frac{\pi(b^2 - a^2)}{2\ln(b/a)} \tag{22}$$

and for the annular thrust bearing as:

$$W = K_g(P_0 - P_a)\frac{\pi(b-a)^2}{\ln(b/a)}. \tag{23}$$

The relationship between the gauge pressure ratio $K_g$ and the clearance $h$ is obtained by equating the flow through the feed holes to the flow in the clearance space. By this means a slot factor $G$ is found which is given by

$$G = \frac{P_a/P_0}{(1-P_a/P_0)^{1/2}(1+P_a/P_0)}\frac{24\mu(2RT)^{1/2}}{P_a}\frac{C_D d^2 \ln(b/a)}{8h^3}$$

and is seen to be very similar to that obtained for journal bearings. Then again the incompressible relationship between $K_g$ and $G$ given in equation (14) can be used to give a relationship between $K_g$ and $h$ similar to that shown in figure 28. At the point of maximum stiffness $K_g = 0{\cdot}69$ and $G = 1{\cdot}25$.

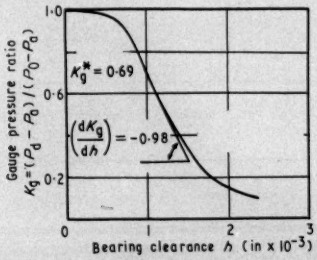



Figure 28. Variation of thrust bearing gauge pressure ratio with clearance, based on incompressible flow theory designed for maximum stiffness at $h^* = 10^{-3}$ in, for other values $h^*$ ($10^{-3}$ in). $dK_g/dh = -0{\cdot}98/h^*$.

In most cases the load capacity of externally-pressurized gas thrust bearings predicted using incompressible flow theory underestimates the measured load capacity by about 10%. The method is therefore suitable for most design purposes as it incorporates a factor of ignorance which operates in favour of the designer.

Shires and Dee (1967) described a design method for narrow annular gas thrust bearings with annular slot feeding and Porritt (1969) has developed the method for application to bearings of all widths. Porritt integrated the compressible pressure distribution in the outward flow region to obtain an expression for load capacity in terms of the error function. He dealt with the inward flow region by expressing the load capacity in terms of the Dawson integral. By these means he obtained an exact solution for load capacity which took account of compressibility and upon which he was able to base an improved approximate design method.

5.1. *Pneumatic hammer*

Both journal and thrust bearings of the externally pressurized type can be prone to an instability known as pneumatic hammer although the problem is usually more frequently met with in thrust bearings. Pneumatic hammer is a self-exciting instability which can occur either with or without rotation. It is almost always audible and in severe cases the bearing surfaces may contact once in each cycle. The onset and degree of severity of the instability is influenced not only by the design of the bearings but also by the stiffness and damping of the structure supporting the machine in which the bearings are employed. The analysis of pneumatic hammer is therefore very complex and no complete theory has been developed. However experience has provided some understanding of the mechanism involved and practical methods have been developed to control its occurrence.

Pneumatic hammer is driven by the energy inherent in the gas flowing through the bearing. Its occurrence appears to be conditional upon the presence of some storage volume within the bearing. The gas film constitutes a finite volume but this is not often of itself sufficient to cause the instability. Bearings with jets feeding into pockets of large volume are, however, very prone to be unstable and the larger the volume the greater is the range of conditions under which the vibration occurs. When a change of load occurs in a thrust bearing the clearance adjusts to a point where the new pressure distribution balances the applied load. However, the rate at which the pressure in the bearing can change is inversely proportional to the volume in the bearing. A large volume means that the pressure adjusts slowly to a change of load and so the pressure change lags behind the change in bearing clearance. If this effect is large enough to nullify the squeeze film damping effect the pressure changes can drive an axial vibration of the rotor.

In practice pneumatic hammer is avoided by designing thrust bearings with shallow grooves rather than pockets. The feed jets supply the grooves which are arranged to outline the desired pocket area. Slot fed bearings do not incorporate pockets or grooves and are consequently less prone to pneumatic hammer. Geometric accuracy is very important in relation to stability and in particular concave bearing surfaces are to be avoided.

If a carefully designed bearing is found to be unstable it can be stabilized by either of two methods or a combination of both methods. The first is by means of a closed cavity which is connected to the pocket or groove by one or more metering orifices. The size of the orifices or the volume of the cavity can be adjusted to tune out the instability. The second method depends upon controlling the exhaust from the bearing. Most machines can be designed so that the exhaust from the thrust bearings flows to atmosphere through a single hole. When instability occurs a metering orifice placed in the hole can be adjusted in size so that the instability ceases while at the same time the load capacity of the bearing is not significantly reduced. By these means pneumatic hammer is controlled and seldom now poses a serious problem.

## 6. Applications of externally pressurized bearings

It is not possible here to describe in any detail the numerous applications to which externally pressurized gas bearings are being applied in machine tools and scientific, industrial and medical instruments. However, an attempt will be made to illustrate with a few examples the advantages which the use of air bearings can confer.

The successful application of externally-pressurized air bearings to precision grinding machines was first reported by Wunsch (1960). It was demonstrated that the eccentricity of spindle rotation was between four and ten times less than that achieved using hydrodynamic oil bearings or rolling contact bearings. Consequently the air bearing enabled substantial improvements to be made to workpiece geometry and surface finish. Cylindrical workpieces rotating on an air-bearing workhead and ground by a wheel rotating on an air-bearing wheelhead can be produced round to within $3 \times 10^{-6}$ in ($7{\cdot}5 \times 10^{-5}$ mm) tir and with a surface finish better than $1 \times 10^{-6}$ in cla.

However, the air bearing does not only produce better quality work but also provides

greater productivity. Grinding machines with conventional bearings are usually left to run for 1–2 h before use on precise work because it is necessary for them to reach a condition of thermal equilibrium. Air bearings used in grinding machine wheelheads do not warm-up and consequently permit the machine to be used immediately after starting. The air-bearing machine is also capable of greater productivity in terms of the rate of metal removal since as the bearings absorb negligible power the full motor power is available to operate on the workpiece.

When grinding tungsten carbide and other ceramic materials it is necessary to use diamond impregnated grinding wheels which are very expensive. The ability of the air bearing to rotate the wheel about a true axis results in even wheel wear and a wheel life that can be extended by up to four times compared to that achieved with conventional bearings.

British manufacturers have taken advantage of the pioneering work undertaken at the National Engineering Laboratory by Wunsch and others and today the majority of small and medium size precision grinding machines which are manufactured in this country are available with air bearings as an alternative to conventional bearings. For several years the Churchill Machine Tool Co. Ltd have offered air-bearing wheelheads on most of their range of grinding machines. Jones and Shipman Ltd will supply their range of surface grinding machines fitted with air-bearing wheelheads manufactured by Westwind Turbines Ltd, and Westwind also make a range of spindles to suit a wide variety of other British and foreign machines. The British Aircraft Corporation (Operating) Ltd and Joseph Lucas Ltd have also been active in developing air-bearing grinding spindles.

Another industrial application to which air bearings have been applied with great success and in which they are now firmly established is the drilling of multi-layer electronic printed-circuit boards. This development was first pioneered by the IBM Corporation at Endicott, NY in 1964/65 using air-bearing drilling spindles manufactured by Barden in the USA and Westwind in England. The technical problem was to drill through the successive layers of copper and fibre board without excessive generation of heat causing smearing of resin over the exposed copper surfaces. The aim was to drill the holes without smearing or burring so that the plating of gold through the holes could follow immediately after the drilling operation without the need for any intermediate time consuming semi-manual operations. It was discovered that the best quality holes, combined with a fast production rate, could be achieved using drills rotating at speeds in the range 80 000–120 000 rpm and driven into the board at feed rates between 2 inches and 4 inches per second. The machines that were finally developed to drill the printed circuit boards of the IBM series 360 computers incorporated 24 air-bearing drilling spindles and were capable of drilling a board with more than 6 000 holes in less than one minute.

Most printed circuit board drilling machine manufacturers now provide air-bearing drilling spindles where high rotational speeds are required, and these machines are widely used throughout the world's electronics industries. The early high-speed air-bearing drilling spindles were driven by means of small air turbines (figure 29). Both the bearings and the turbine received compressed air from the same source and the units were simple in design and installation. However, on multi-spindle machines the demand for compressed air can become excessive and it is more economical to use small high frequency electric motors. Westwind Turbines Ltd have developed two electrically-driven air-bearing drilling

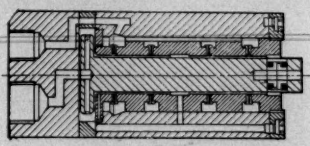


Figure 29. 120 000 rpm air turbine drilling spindle with air bearings (by courtesy of Westwind Turbines Ltd).

spindles that are capable of speeds up to 100 000 rpm and with powers of 100 W and 300 W. Since introducing them in 1965 Westwind have supplied more than 2 000 air-bearing drilling spindles of both air turbine and electric motor types for application to electronic printed circuit boards.

Externally pressurized air bearings have found many applications in scientific and industrial instruments. Their extreme accuracy of rotation has prompted their use in roundness measuring machines and one machine developed at the National Engineering Laboratory is now produced by Optical Measuring Tools Ltd. Another machine was developed by the British Aircraft Corporation (Operating) Ltd. These machines are capable of detecting deviations from a true circle down to a limit of $1 \times 10^{-6}$ in ($2{\cdot}5 \times 10^{-5}$ mm).

The low friction of air bearings has prompted their use in many instruments. Two of the earliest were the wind tunnel balance developed at the Royal Aircraft Establishment and the precision dynamometer described by Graneek *et al.* (1959). They have also been applied to goneoviscometers, pulleys for fibre and thin film tensometers and rollers for flying micrometers used in the control of the rolling of thin metal plate. A particularly successful application has been in turbine flowmeters for gases. By freely mounting a turbine on a low friction bearing the speed of the turbine is linearly related to the velocity of the gas flowing past the blading. The gas bearing turbine flowmeter was developed at the University of Southampton in 1960/61 and instruments of this type are now produced by Aerostatic Ltd and by Westwind Turbines Ltd. The aerostatic design (figure 30) is notable

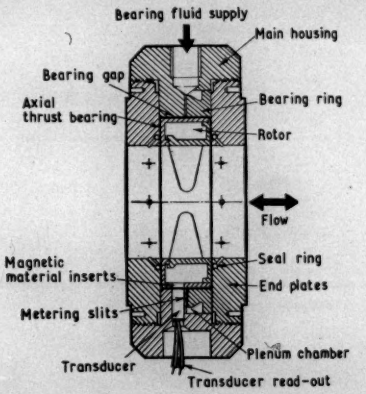



Figure 30. Turbine flowmeter with slot fed air journal bearing (by courtesy of Aerostatic Ltd).

for its use of slot fed bearings which permits its application to any gas and also to many liquids without any adjustment of the bearing dimensions. The Westwind design employs a small jet fed journal bearing located through a central cylinder around which are mounted the turbine blades. This arrangement minimizes the friction torque of the bearing and extends the linear range of the instrument. It has been employed by ICI Plastics Ltd for monitoring and controlling the flow of process gases in the manufacture of plastics and has demonstrated an accuracy of $\pm 1\%$.

There are many applications to instruments which employ air bearings operating at high speeds. Montgomery and Steiry (1956) reported on a bearing system for a high-speed camera capable of speeds of up to 250 000 rpm. They used bearing bushes made of

porous bronze and the compressed air was fed through the porosity to the bearing clearance. Rubber rings encircled the bushes to seal the air supply and to suppress self-excited whirl. The system was later employed in a range of neutron choppers which are still in use at AERE Harwell.

During the last decade a number of instances have arisen when physicists have required high-speed rotors capable of carrying a mirror, prism or chopping disk for use in infrared or optical choppers or laser 'Q' spoilers. Several of these devices have been developed by Westwind Turbines Ltd for the National Physical Laboratory. These have included units adapted from industrial products as well as specially designed units. In the former category is a modified hand tool which rotates a small chopping disk at speeds of up to 600 000 rpm and a modified air turbine grinding spindle which rotates a 2-in glass prism at speeds up to 30 000 rpm. In each case essential features are precise rotation and a very low level of vibration.

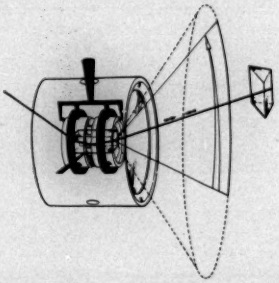


Figure 31. Optical scanning turbo-prism (by courtesy of the National Physical Laboratory).

Figure 31 shows a diagram of a laser 'Q' switch which was specially developed to meet an NPL requirement. The engineering development of this unit has been described by Powell (1969) and its application has been reported by Gates *et al.* (1968). The rotor takes the form of a hollow cylinder in the bore of which is located a pentagonal prism with internally reflecting surfaces. The drive is provided by a small turbine of the Pelton Wheel type which is centrally located around the outside of the rotor. The two air journal bearings occupy the remainder of the cylindrical surface of the rotor and narrow end faces provide axial location. The design of the rotor was adapted so that light incident upon the end faces of the prism within a solid angle of 120° could pass through unimpeded. The earliest model of the turbo prism 'Q' switch rotates a prism of 10 mm aperture at a speed of 90 000 rpm. An improved version with the same size of prism runs at 130 000 rpm and the latest development carries a prism of 4 mm aperture at speeds in excess of 250 000 rpm.

The development of the technology of gas lubrication has owed much to the science of physics and it is fitting to record that in some small measure the debt is being repaid.

**References**

AUSMAN, J. S., 1957 a, *Trans. ASME*, **79**, 1218.
—— 1957 b, Proc. Conf. on Lubrication and Wear.
—— 1961, *J. basic Engng*, **83**, 188.
BURGDORFER, A., 1959, *J. basic Engng*, **81**, 94.
DRESCHER, H., 1959, *Proc. 1st Int. Symp. on Gas Lubricated Bearings*, Washington D.C., pp. 319–45.
DUDGEON, E. H., and LOWE, I. R. G., 1965, *Mech. Engng Rep. natn. Res. Coun. Can.*, MT-54.

FISCHER, G. K., CHERUBIM, J. L., and FULLER, D. D., 1959, *Trans. ASME*, **58A**, 231.
FORD, G. W. K., HARRIS, D. M., and PANTALL, D., 1957, *Proc. Instn. Mech. Engrs*, **171**, 93.
GATES, J. W. C., HALL, R. G. N., and ROSS, I. N., 1968, *Proc. 8th Int. Cong. on High Speed Photography*.
GRANEEK, M., WUNCH, H. L., and NIMMO, W. M., 1959, *Engineering*, **188**, 4886, 614.
GROSS, W. A., 1962, *Gas Film Lubrication* (New York: Wiley).
—— 1967, *Proc. Gas Bearing Symp.*, Southampton Univ. Paper 23.
GROSS, W. A., *et al.*, 1959, *IBM J. Res. Dev.*, **3**, 237.
HARRISON, W. J., 1913, *Trans. Camb. phil. Soc.*, **22**, 39.
KATTO, Y., and SODA, N., 1961, *Trans. ASME* No. 61, Lub. 8.
KERR, J., 1962, *Proc. Instn. Mech. Engrs*, **176**, 582.
LICHT, L., 1969, Minutes of ONR–UK Gas Brg. Res. Tech. Coord. Meeting, 20.
LUND, J. W., and VOHR, J. H., 1967, *Proc. Gas Bearing Symp.*, Southampton Univ. Paper 29.
MALANOSKI, S. B., and PAN, C. H. T., 1964, *Mech. Tech. Inc. Rep.* MT1-64TR11.
MARSH, H., 1965, *Instn. Mech. Engrs, Mech. Engng Sci.* Monograph No. 2.
MECH, C., 1969, *Proc. Gas Bearing Symp.*, Southampton Univ. Paper 4.
MONTGOMERY, A. G., and STERRY, F., 1956, *AERE Rep.* No. E.D./R.1671.
MUIJDERMAN, E. A., 1966, Philips Technical Library (Cleaver-Hume Press).
NG, C. W., and PAN, C. H. T., 1965, *J. basic Engng*, **87**, 675.
PANTALL, D., 1964, *Gas Lubricated Bearings*, Eds N. S. Grassam and J. W. Powell (London: Butterworth's), pp. 27–84.
PORRITT, T. E., 1969, *Proc. Gas Bearing Symp.*, Southampton Univ., Paper 23.
POWELL, J. W., 1963, *I. Mech. E. Lub. and Wear Group Conv.* Paper 11.
—— 1964, *Trans. ASME*, 64-WA/LUB-11.
—— 1968, *Trans. ASME*, 68-LUB-9.
—— 1969, *Proc. Gas Bearing Symp.*, Southampton Univ., Paper 2.
—— 1970, *Design of Aerostatic Bearings* (London: Machinery Publishing Co. Ltd).
POWELL, J. W., and TULLY, N., 1962, *Engineer*, Lond., **213**, 1121.
RAIMONDI, A. A., 1961, *Trans. ASME* 4, No. 1, 131.
ROBINSON, C. H., and STERRY, F., 1958, *AERE Rep.* No. R/R 2642.
SHIRES, G. L., 1949, *NGTE Rep.* No. R61.
SHIRES, G. L., and PANTALL, D., 1963, *I. Mech. E. Lub. and Wear Group Conv.* Paper 9.
SHIRES, G. L., and DEE, W. C., 1967, *Proc. Gas Bearing Symp.*, Southampton Univ., Paper 7.
SIXSMITH, H., 1959, *Proc. 1st Int. Symp. Gas Lubricated Bearings*, Washington D.C., 418.
TULLY, N., 1966, *Ph.D. Thesis*, Southampton University.
WHITLEY, S., 1964, *Gas Lubricated Bearings*, Eds N. S. Grassam and J. W. Powell (London: Butterworth's), pp. 85–103.
—— 1967, *Proc. Gas Bearing Symp.*, Southampton Univ., Paper 13.
WHITLEY, S., and BETTS, C., 1959, *Br. J. appl. Phys.*, **10**, 455.
WHITLEY, S., and WILLIAMS, L. G., 1959, *UKAEA I.G. Rep.* 28 RD/CA.
WHITLEY, S., BOWHILL, A. J., and MCEWAN, P., 1962, *Instn. Mech. Engrs*, **176**, 554.
WHIPPLE, R. T. P., 1958, *AERE Tech. Rep.* 622.
WILDMAN, M., 1957, *M.Sc. Thesis*, Univ. College of Los Angeles.
WOOD, M. D., 1968, *I. Mech. E., J. mech. Engng Sci.*, **10**, 4.
WUNSCH, H. L., 1960, *Proc. 1st Conf. on Machine Tool Design and Research*, Birmingham.

C. L. BOLTZ

# Nobody Breathes Fresh Air

## Facts and Arguments about Air Pollution

Man is by his very existence a polluter of his environment, even if he remains in bed from birth to death. He breathes in 250 ml of oxygen every minute and breathes out 200 ml of carbon dioxide. Had it not been that plants in daylight consume carbon dioxide and emit oxygen in the process of photosynthesis mankind would have been gassed out of existence long before now. As it is, man has multiplied and is still multiplying in numbers and he has reduced the area producing green plants. Furthermore by the increasing success of his technology he has increased the $CO_2$ pollution to the point where that hardy annual of newspaper stories, the melting of the polar icecaps, gains credibility to the extent that the Government has called for the opinion of its newly-established Royal Commission (on pollution of the environment) on the adequacy of present research on this matter.

This restating of the forgotten obvious may serve to keep the ensuing discussion cool. Air pollution is apt to get among the topics of the inveterate cause-supporter and the person who wishes to climb on the current anti-technology bandwagon. It is then distorted; graphical and statistical extrapolations are made which no unbiased scientist would accept; the fact that air pollution is an accompaniment of living is ignored. The problem to be discussed, therefore, is not that of the complete elimination of all pollution but that of controlling it so there shall be a minimum of interference with the natural flora and fauna, of controlling it in the interest of health and amenity (even smells reduce the quality of living) and the prevention of damage to materials and structures. Only the last of these can be quantified to any usable extent, though data about the loss of working efficiency, such as the days lost because of chronic bronchitis (39 million days in Britain in the year 1967–68), can contribute something to the assessment of pollution hazards.

The questions about the pollution of the gaseous part of the biosphere which the physicist and technologist must try to answer include: What are the pollutants of the atmosphere? How much is there of each? What is the variation in time and space? How can they be detected and measured? How are they caused? How can they be reduced? If amelioration is to be achieved how much will it cost? The questions about who is to pay and what steps are to be taken to achieve control are answerable only by central and local government. The questions about the effect of the pollutants on health must be answered by medical research.

Without any of man's technology and industry the atmosphere, ideally composed of oxygen, nitrogen, carbon dioxide, argon and traces of other gases, and water (in sizes that range from the molecule to the droplet), according to elementary textbooks, in reality contains nitrogen oxides, carbon oxides, hydrocarbons, dust, salt, sulphates, ammonia and so on due to natural pollution. Meteoric dust, volcanic eruptions and dust storms, etc. add particulate matter. An interesting example was reported from Harwell last year, when particles of regular geometrical appearance, rather like radiolarian skeletons, were seen in atmospheric samples under high magnification. They were about 0·5 $\mu$m in diameter and help was sought through *Nature* for their identification. They were eventually shown to be insect droppings. According to one authority pigeons contribute significantly to airborne particulate matter in New York's atmosphere. Nitrogen compounds come from biological reactions; sulphates, salt, lead, etc. come from sea spray, hydrocarbons from vegetation. The ultraviolet produces ozone which reacts with hydrocarbons to produce more complex compounds. Thus it can be seen that a man believing that he is breathing pure fresh air is in fact taking into his lungs a whole range of substances that are pollutants.

**Table 1. Main forms of air pollution in the United Kingdom**

| | Gas molecules | Very fine particles | Visible particles |
|---|---|---|---|
| Widespread pollutants | Sulphur dioxide (Measured daily for the National Survey) | Smoke (Measured daily for the National Survey) | Grit and dust (Measured monthly) |
| Typical size | 1 nm | 1 $\mu$m | 100 $\mu$m |
| Properties | Mix with air<br>Invisible<br>Corrosive | Float with air<br>Individually invisible†<br>Soil all surfaces | Fall from air<br>Visible<br>Soil where deposited |
| Health | Irritant<br>Poisonous if strong | Carried to lungs in breathing | Trapped in nose and throat |
| Sources | Sulphur in fuels (all types of appliance) | Inefficient burning (mainly domestic) | Entrained ash, soot and fuel dust (mainly industry) |
| Other pollutants of local importance | Oxides of nitrogen<br>Carbon monoxide in dense stagnant traffic | Sulphuric acid aerosol (mist)<br>Diesel smoke<br>Industrial 'fume' | Industrial dusts (from manufacture or storage) |

† Once a smoke plume is diluted, the particles are only perceived in certain lighting conditions, either as a haze, or in a beam of light.

Figures for the quantities of these natural pollutants as well as for man-made pollutants can be confusing and misleading. For instance, some $10^8$ tons of natural particulate matter are added to the atmosphere every year. But the whole atmosphere weighs $5{\cdot}7 \times 10^{15}$ tons. On the assumption of perfect diffusion, therefore, the dilution is so small as to be inexpressible even as a percentage. Thus the figure of such an immense number of tons seems impressive until some arithmetic is done. It has also to be remembered, however, that most man-made pollutants are not perfectly distributed but exist in small localities in much higher concentrations than the tonnage figures would suggest when expressed in relation to the whole atmosphere A further confusion is that concentrations are often expressed in different units, such as parts per million, percentage, or weight per unit volume The growing British method of using microgrammes per cubic metre ($\mu g\ m^{-3}$) allows gases, liquids and solids to be expressed in the same way.

Before going further into the discussion we can determine what the man-made pollutants are. Those most frequently measured are given in the Warren Spring Laboratory table 1. $CO_2$ is not mentioned there because it is not normally considered a pollutant, though it can kill by asphyxiation as it has done in sunken submarines. There are hydrocarbons, including at least one that is a carcinogen. There are others confined to special works and affecting only people working in them—e.g. chlorine, benzene, isocyanates, chromic acid. These are not usually included in discussions of atmospheric pollution because of their confinement, though small amounts can get out into the atmosphere. There are compounds that are created in the atmosphere with the help of ultraviolet radiation and make nuclei for such aerosols as the photochemical eye-smarting smog that envelopes Los Angeles like a gauze. The effect of hydrocarbons is to produce sticky, tarry material that adheres to surfaces with its entrained solid or gaseous pollutant. Tarry deposits and grit can alight on foliage and interfere with its respiration and transpiration. There are fluorides that destroy vegetation. Airborne pollutants can fall into water and interfere with aquatic flora and fauna. Grit settles on and in homes. In all cases, it must be remembered, the concern is with pollutants near the ground where man lives, with concentrations that may be very local.

There is some conflict of evidence and difference of opinion about the effects of pollutants on health. Dr Patrick Lawther, Director of the Medical Research Council Air Pollution Unit at St Bartholomew's Hospital Medical College, has pointed out that moderate cigarette smokers have much more carbon monoxide in their blood than people subjected for

hours to London's traffic. On the other hand the Consumers' Association in a report in *Which?* in November last claimed that tests of mental performance given to four young men who breathed air from 15 inches above the kerb showed significant fall-off in performance. It gave no detailed results of the tests or the measurement of significance and Dr Lawther and his team are now pursuing a number of investigations into the possible effects of CO—in concentrations found in real life—on man. Furthermore some workers claim that CO is not a cumulative poison in small doses whereas others claim that it may well be. The effects of $SO_2$ are in doubt. The several thousand deaths in London owing to the smog of 1952 was at first attributed to $SO_2$ but in a similar smog in 1962 only 700 died. Between the two events the Clean Air Act of 1956 had come into force, reducing the amount of dark smoke. Yet nobody would suggest that $SO_2$ is a good thing to breathe, and as for CO, the evidence of accidental deaths and suicides shows how toxic it is.

The truth is that the human body is a very adaptable biochemical machine which can withstand much insult but this does not mean that the insult should be increased or that we should encourage it or put up with it. Considerations such as these on the medical side together with the multivariate nature of man's metabolism make it very difficult to get statistical evidence that points to unexceptionable conclusions. The correct medical research must therefore go on, as it does in the MRC's unit mentioned, in the Ministry of Health, the College of General Practitioners, the London School of Hygiene and Tropical Medicine, St Thomas's Hospital Medical School, the West Regional Hospital Board's laboratories in Scotland, and the Welsh School of Medicine. These are the main British centres; there are many more in other countries, especially the USA.

There are other aspects of pollution that are the subject of continuing investigation overseas and in Britain. The Ministry of Agriculture, Fisheries and Food, in its laboratories at Weybridge, London and Hoddesdon, investigates the effects of $SO_2$ and soot on greenhouses and greenhouse plants, the contamination of herbiage by particulate and gaseous fluorine compounds and the husbandry practices to alleviate the effects of air pollution. The Building Research Station is interested in the effects on the durability of building materials. The University of Oxford is interested in the genetics of colour in certain animals such as moths that have bred black in the north country.

The political aspect of air pollution has been the subject of attention for a long time. It is historically entertaining to learn that Edward I banned coal in London in 1306 because he considered the smoke unhealthy, but even when one looks at more modern times one finds that there has been concern about atmospheric pollution for a great many years and indeed the penalties for some pollutions are said to go back to 1804. The Alkali Acts, deriving from the fact that alkalis were the foundation of the British chemical industry, have been concerned, the latest being the Alkali Act of 1906. It is an interesting comment on the odd way that British public affairs grow on historical beginnings that the Alkali Inspectorate is still responsible for overseeing the emissions into the atmosphere of 'noxious' gases from chemical factories, iron and steel works, gasworks and so on, including power stations and such establishments as works for processing uranium and other metals and works for burning off PVC on scrap cable. The Factory Acts have been responsible for pollution conditions inside factories for many years and have established criteria for acceptable levels with the help of the Laboratory of the Government Chemist. Smoke abatement was mentioned in an Act of 1947 and the Public Health Act of 1875. Thus it can be seen that public concern is not especially a modern phenomenon.

The wider concern today may be attributed to several factors: (i) the *increase* in population, urbanization and sources of pollution (industry and the motor car); (ii) the changing social conditions in which people wish to live without stench and dirt and smoke—they are not so concerned with the less obvious pollutants; (iii) the unflagging efforts of organizations such as the National Society for Clean Air in Britain and similar organizations abroad, leading to the establishment of the International Union of Air Pollution Prevention Associations, which held its first congress in London in 1966; (iv) the influence of mass media such as radio and television; and (v) the death of 4000 people in the London smog of 1952.

The last-mentioned had immediate effects. The Beaver Committee was formed and reported in 1954, whereupon Sir Gerald Nabarro introduced a Clean Air Bill and withdrew it on the Government's promise to introduce an official Bill. This led to the Clean Air Act of 1956, which was concerned only with dark smoke (shade 2 or worse on the Ringelmann Chart) and grit and dust, all from chimneys (domestic fires were worse offenders than industrial combustion). The Act made it an offence to issue such smoke or fail to take steps to minimize the output of grit and dust. There were exemptions (e.g. Government establishments!) and permitted delays and appeals, much of which left loopholes soon seen by the wily law-evader. It also gave local authorities the right to establish 'smoke control areas'. Regulations to supplement the Act followed, such as how to get the correct height of chimney, the time allowed for the emission of dark smoke and black smoke (shade 4 Ringelmann) and so on. Then the Clean Air Act of 1968 tidied up the previous one by widening the range of furnaces covered, extending the possible list of offenders by not restricting the emissions to chimneys and empowering the Ministry to *order* a local authority to declare a smoke control area. It also forbad anyone to buy or sell in such areas solid fuel, such as bituminous coal, that would produce smoke.

This last enactment ironically worsened the position instead of improving it because there is not enough smokeless solid fuel and there will not be in the winter of 1970–71. Lord Robens has suggested that the Government should not authorize any more smoke control areas. The absence of a carefully-thought-out fuel policy seems to blame in the last analysis. The result is that at the moment of writing 15 local authorities have succeeded in getting their scheduling as smoke control areas suspended.

It is worth recording that the City of London became a smokeless zone in 1955, before the Act was on the Statute Book, and that the then Metropolitan Borough of Holborn was in 1962 the first local authority in Britain to reach its smoke control area target. Over 350 local authorities have already become, or are in process of becoming, smoke control areas.

Numerous figures and tables have been produced to show how much pollution is being put into the atmosphere of Britain. Dr Albert Parker, well known in world-power circles, has estimated the amount of smoke, $SO_2$, CO and grit and dust put into the air. (Actual measured concentrations of the first two are provided by the National Survey to be mentioned later. These figures in tables 2–5 are emissions.) Table 2 shows his estimates for 1938, 1960 and 1968 for smoke and $SO_2$ from coal and oil.

Dr Parker's estimates for grit and dust are, he admits, very rough, partly due to difficulty of measurement and partly to the few places where it is done.

The figures were provided by the National Society for Clean Air. They are not easy to collect and collate because of the different sources used, each with different criteria of classification, and Dr Parker is one of the few masters of this expertise, so that his estimates can be taken seriously.

The first thing to be realized from the figures is that these impressive millions of tonnes must be taken in relation to the mass of the atmosphere. How much diffusion is there? Radioactive particles from bomb tests have been found high up, which would have been expected, and have eventually reached the ground. Particles have also been found, for example, at 70 000 m and collected by rocket and examined by electron microscope and x-ray analysis and proved not to be micrometeorites. This sort of fundamental research on far-dispersed pollutant is sometimes relevant to what is happening at ground level and further research may produce a usable relationship. Furthermore, widespread dispersion is a way of reducing pollution that affects us, a point to be remembered. However, it is concentration in our immediate environment that is important and as far as health is concerned it is the threshold that produces clinical symptoms that is important. Local conditions may produce concentrations above threshold values. Los Angeles, for example, may be considered to some extent as a closed system. The weather, topography, and localization of industry are all important. The canyons called streets in some cities are thus important in the concentration of vehicle pollutants.

**Table 2. Smoke and sulphur dioxide in $10^6$ metric tonnes**

| | 1938 | 1960 | 1968 |
|---|---|---|---|
| Smoke | | | |
| From *coal* | | | |
| Domestic | 1·74 | 1·21 | 0·75 |
| Railways | 0·26 | 0·07 | small |
| Industrial and miscellaneous | 0·75 | 0·19 | 0·09 |
| Total | 2·75 | 1·47 | 0·84 |
| Sulphur dioxide | | | |
| From *coal* | | | |
| Domestic | 1·23 | 0·92 | 0·60 |
| Electricity power stations | 0·41 | 1·56 | 2·11 |
| Railways | 0·37 | 0·26 | 0·01 |
| Collieries | 0·32 | 0·14 | 0·07 |
| Industrial and miscellaneous | 1·35 | 1·22 | 0·77 |
| Coke ovens | 0·07 | 0·11 | 0·09 |
| Gas industry | 0·14 | 0·16 | 0·07 |
| Total | 3·89 | 4·37 | 3·72 |
| From *oil* | | | |
| Domestic | — | small | 0·01 |
| Industrial and commercial | 0·05 | 1·17 | 2·01 |
| Road and rail transport | 0·01 | 0·05 | 0·09 |
| Marine craft (inland) | — | 0·04 | 0·02 |
| Total | 0·06 | 1·26 | 2·13 |
| From *coke* | | | |
| Domestic and smokeless fuels | 0·06 | 0·10 | 0·14 |
| Industrial | 0·18 | 0·26 | 0·15 |
| Total | 0·24 | 0·36 | 0·29 |
| Total $SO_2$ | 4·19 | 5·99 | 6·14 |

**Table 3. Grit and dust in $10^6$ metric tonnes**

| | 1938 | 1960 | 1968 |
|---|---|---|---|
| Domestic fires | 0·145 | 0·1 | 0·10 |
| Power stations | 0·145 | 0·4 | 0·32 |
| Railways | 0·145 | 0·1 | — |
| Coal and coke used industrially | 0·215 | 0·2 | 0·18 |
| | 0·65 | 0·8 | 0·6 |
| Add grit and dust from various industrial processes | 0·5 | 0·5 | 0·5 |
| Total | 1·15 | 1·3 | 1·1 |

(Figures are decimalized from Dr Parker's rough fractions.)

**Table 4. Carbon monoxide in $10^6$ metric tonnes**

| | 1968 |
|---|---|
| From | |
| Industry | 6·0 |
| Domestic | 5·0 |
| Petrol engines | 6·1 |
| Diesel engines | 0·1 |
| Total | 17·2 |

**Table 5. Pollutants from road vehicles in $10^6$ metric tonnes**

| | 1968 | |
|---|---|---|
| | Petrol engines | Diesel engines |
| Carbon monoxide | 6·1 | 0·10 |
| Hydrocarbons | 0·30 | 0·18 |
| Aldehydes | 0·01 | 0·003 |
| Nitrogen oxides | 0·21 | 0·05 |
| Sulphur dioxide | 0·02 | 0·03 |
| Total | 6·64 | 0·363 |

*Note:* Figures for CO were also given in table 4.

A few inferences can be made from the data of tables 2–5. There is for example the trend over the years, and this will be considered later. There is the small variation in grit and dust, a surprising fact, from 1938 to 1968. As far as CO is concerned domestic fires add almost as much as industry and as much as petrol engines, and these are more offensive than diesel engines.

The world situation is little different from that in Britain. From tables 2–5 the total mass of pollution for 1968 is seen to be $25{\cdot}28 \times 10^6$ tonnes. That for the USA for 1966 (given by Professor P. S. Myers in a paper read to the Institution of Mechanical Engineers in November 1969) was $142 \times 10^6$ tons (whether long or short not stated, a further example of the difficulty in the assessment of data). Dividing these figures by the populations, taken as 50m for the UK and 200m for the USA, one gets the *per capita* pollution for the two countries as 0·51 tonne for the UK and 0·71 tons for the USA. These are of the same order and the US higher figure is due to more pollution by road vehicles, a fact that has led to much concern in the USA and caused President Nixon to consult his Environmental Control Council. Car manufacturers have co-operated—being compelled to in Los Angeles—but it must be remembered that powerful American cars can be cut down slightly in speeds available without much affecting the American way of life. The close comparison between the figures for the UK and the USA can be extended to other advanced and crowded countries. The research and the control methods are thus on the whole universal and the problems being attacked by physicists, chemists and technologists in Britain can be taken as sufficiently representative of the world situation to justify no attempt here to assess world-wide research.

One can sum up the world situation as in table 6 for the pollution caused in the world as a whole (W) or in the annulus between the 30th and 60th parallels of latitude (P), the region in which most advanced industrial countries exist. The figures are taken from Professor Myers.

**Table 6. Total pollutants in $10^6$ tons**

| Pollutant | Natural | Man-made | W = whole P = part |
|---|---|---|---|
| Carbon dioxide | $2 \times 10^5$ | $2 \times 10^4$ | W |
| Carbon monoxide | $2 \times 10^2$ | $2 \times 10^2$ | W |
| Sulphur dioxide | negligible | $1 \times 10^2$ | P |
| Nitrogen oxides | $2 \times 10^2$ | 50 | P |
| Hydrocarbons | $3 \times 10^2$ | 90 | W |
| Particulates | $10^2$ | $10^2$ | P |

From this table it can be seen that the only important pollutant that is entirely man-made is $SO_2$ and that for no other pollutant does that made by man exceed that from natural occurrences. Indeed in some cases the man-made quantity is much less than the natural quantity. Taking the mass of the atmosphere as $5{\cdot}7 \times 10^{15}$ short tons the proportion of pollutants is very small indeed—$1 \times 10^{-7}$ for CO. Once again it must be emphasized, however, as it has been emphasized before, that complete diffusion cannot be assumed and that local conditions are much more important. Round Los Angeles, for example, the concentration of nitrogen

oxides is 100 times as high as the world average and in industrial sections of cities where much solid fuel is used concentrations of $SO_2$ are thousands of times that in the atmosphere as a whole. Carbon monoxide in moving vehicles has been measured in concentrations enough to cause headache and reduced mental acuity whereas most concentrations in urban areas would have no effect at all.

Facts such as these make careful and widespread surveys necessary. Regional variations in Britain are shown in recent analyses (8 June 1970) made by the Warren Spring Laboratory. The average concentration of smoke and sulphur dioxide in the northern region in the past year, in $\mu g\ m^{-3}$, is 108 and 97 respectively, whereas for Greater London it is 46 and 151, showing much less smoke in London but much more $SO_2$. The corresponding consumption of domestic coal per head is 0·69 tonne and 0·06 tonne. These figures show how important it is to have detailed and widespread data. The only inference seems to be that people in the north like their coal fires!

**Table 7**

| Pollutant | Source |
|---|---|
| $CO_2$ | All combustion and respiration. |
| CO | Combustion of all sorts—high correlation coefficients have been obtained between traffic density and CO emission in selected parts of the USA. |
| $NO_x$ | Biological reactions, power stations, road vehicles, space heating, waste disposal. |
| $CH_4$ and other hydrocarbons | Natural events, industry, oil combustion, road vehicles, incinerators. |
| $SO_2$ | Industry, power stations, space heating, combustion of coal and oil, road vehicles. |
| Grit and dust | Natural dust, industrial dust, ash from combustion, road vehicles. |
| Smoke | Combustion of coal, oil and wood, road vehicles, aero-engines at airfields. |

When one considers the atmosphere as a mass of gas in constant turbulent motion one finds that pollutants do not merely accumulate; there is not only gain but also loss. Particulate matter drops out and falls in rain or other precipitation. (This fact has led to an interesting expedition activity now going on in Antarctica, where Dr D. A. Peel of Bristol University is leading a team to examine ice strata (frozen snowfalls) for evidence of pollution from today backwards through the centuries and millenia. There will be on-the-spot chromatographic analysis and snow will be brought back for tests.) Carbon dioxide is removed by photosynthesis and extensive absorption in the oceans. Some carbon monoxide changes to carbon dioxide and some natural events are believed to destroy it though little is known about this. Sulphur dioxide is the one entirely man-made sulphur compound that pollutes; other sulphur compounds such as $H_2S$ and sulphates come from natural sources. $SO_2$ can be oxidized to $SO_3$ and so form sulphuric acid in water and fall in rain. NO is oxidized to $NO_2$ and then further changed to nitrates which fall on the earth in rain and snow. These nitrogen oxides can also react with other pollutants. So does ammonia which is a natural pollutant and sometimes a man-made one. Methane is the most plentiful part of the hydrocarbon pollution, coming it is believed from bacterial decomposition. It seems to stay fairly unchanged but some hydrocarbons react with other compounds and precipitation removes some from the atmosphere. Rain and other precipitation is thus seen to be a great cleanser of the atmosphere, a great remover of pollution, though the water and earth pollution may increase as a result.

The sources of the most-investigated pollutants have been mentioned but it is worth tabulating them as a summary. This is done in table 7.

Attempts have been made to calculate the cost of air pollution. Results were given for Britain in the appendix to the Beaver Report. Direct costs in laundry, painting and decorating, cleaning and depreciation of buildings other than houses, corrosion of metals and damage to textiles and other goods were estimated at £150m a year or thereabouts. Costs due to loss in working efficiency in agriculture, manufacturing industry, transport, etc were estimated at £100m a year The total was £250m a year, made before devaluation, equivalent to about £10 per head per annum in the black areas and £5 per head elsewhere. These figures did not include loss of fuel because of incomplete combustion and they were considered conservative estimates Loss of amenity of course cannot be calculated. Nor can a great deal of damage to health. In the USA a figure of between 7·5 and 11 thousand million dollars has been given for 1962 by the US Department of Health, Education and Welfare. A mean figure of 10 000 million dollars per annum can thus be guessed at for the whole world.

Figures such as these may convince even hard-headed business men insensitive to claims about amenity and health and general enjoyment of the environment that there is clearly a need for the reduction of air pollution. This means research. The non-medical research can be divided into three parts: (i) the detection, measurement and statistical recording and assessment of each type of pollution; (ii) the reduction of pollution before and after emission into the atmosphere; (iii) fundamental investigations of all airborne materials and their mixing and interaction. All of these are the subject of research in many countries.

In Britain there has been some research for a very long time and there were measuring sites for the Meteorological Office as long ago as 1921, and this Office is of course still interested. On the whole the three establishments most concerned with measurement are: (i) the Warren Spring Laboratory (Ministry of Technology), the most authoritative on all matters concerned with air pollution; (ii) the Central Electricity Generating Board with its main laboratory (CERL) at Leatherhead, primarily concerned with power-station emissions; (iii) the UK Atomic Energy Authority at the Harwell establishment, concerned with sophisticated analysis. Harwell came into the air-pollution field in general with its diversification policy from about 1967 onwards, though it had always been concerned with pollution by radioactive materials. The National Physical Laboratory has done considerable work on chimneys and pollutant dispersal with the aid of wind tunnels. The Warren Spring Laboratory has also done and is still doing work on chimneys and dispersal aerodynamics. The Laboratory of the Government Chemist has devised methods for measuring toxic substances in industrial premises and has established with the Factory Inspectorate threshold limit values. Publicly-owned industries concerned in some research, however specialized, are the National Coal Board, the London Transport Executive, the Gas Council and the British Railways Board. Ten of the co-operative research associations have programmes on air pollution; without making invidious comparisons the Motor Industry Research Association should be mentioned separately because of its work on road vehicle pollution. Twelve universities or similar institutions have research programmes. None of all this research is concerned with the *effects* of pollution. There is of course widespread research activity in the USA and in the advanced countries of Europe. Special mention is necessary of the unique National Survey of Air Pollution conducted by the Warren Spring Laboratory. Readings of smoke and $SO_2$ are taken at 1200 sites daily and the sort of equipment used has been accepted by OECD for recommended use in Europe.

The instruments and techniques developed so far show that there is still scope for the inventive physicist in the air pollution field, for not one device is without its disadvantages. The more fundamental analyses use very modern instruments—chromatographs, x ray microanalysers, electron microscopes, lasers, etc.—but the need is for simple, reliable, sensitive, specific and inexpensive devices, either automatic or usable by relatively unskilled personnel. Some methods still used are based on very old notions and are none the worse for having been devised by physicists of the old-fashioned sort who were extremely talented in experimentation and basic in their approach. Who would think, for instance, that the Ringelmann Chart would still be used to judge the darkness of smoke? This is a chart with

five 4-inch squares drawn on it, one empty and the other four with lines drawn at right-angles across the square, the lines being thin for shade No. 1, thicker for shade No. 2 and so on. Seen at a distance of 50 feet the cross lines merge and each square is a tint of grey. Visual comparison of this with a smoke plume determines the Ringelmann number. The judgment depends on the brightness of the sky and is of course subjective, and it is meaningless when used with other than black smoke. However, its acceptance makes it usable—a negotiable entity turns subjectivity into objectivity—and it is simple to the point of childishness.

At first thinking it would seem that the measurement of dust and grit is the easiest of all. It is not. Heavier particles drop more freely whereas finer dust particles produced for example by the combustion of pulverized fuel do not fall so freely and are carried nearly horizontally by the wind. The old system, still in use, of measuring what falls in a deposit gauge is thus not usable in all circumstances. Another system is to catch particles on a sticky surface. Obviously it would be easy to count particles electronically, but the variation in size and concentration and ratio makes simple measurement not so easy at all. OECD in 1965 considered that the effects of dust and grit were very local, which is true, and so there was no incentive to set up an international standard procedure. This has not prevented D. H. Lucas at the Central Electricity Research Laboratory from developing a directional deposit gauge which has been welcomed by the Chief Alkali Inspector. Four gauges are on a vertical support to help in giving an indication of wind direction. The dust (and rain) go through a vertical open space and fall into the bottom. This is removed after, say, a month and the dust and rain poured into a beaker. The optical obscuration is then measured.

It is preferable to monitor dust and grit in flues before it is emitted. The British Coal Utilization Research Association has developed a cyclone probe device for this purpose and it is now available from Airflow Developments. The flue gas is drawn through a tube and the solid matter is thrown to the periphery by centrifugal force. It then emerges through an orifice and can be collected and weighed. This instrument is included in a British Standard and has been specified in one of the Clean Air regulations. For the measurement of submicron particles a glass-wool filter is added to the airflow tube after the cyclone. Needless to say, an experienced person is needed to use this device if the results are to be useful and reliable. CERL has also developed a dust monitor invented by D. H. Lucas and a colleague. It is of course primarily for power stations but could be used for some other industries. Both versions, Mark I and Mark II, the latter needing less maintenance, have been installed in a number of power stations. The Lucas monitor depends on the obscuration by dust or grit of a small glass plate at the bottom of a cylindrical cell. Light passes through this plate and the obscuring changes the photoconductor resistance which is part of a Wheatstone bridge. These changes are recorded on a chart. The object, of course, is to get routine evidence of the failure of the system being used to eliminate dust and grit before emission. An older type of instrument is an optical density meter for smoke and fine dust. This type is also still in use and results have to be correlated with the Ringelmann number. This has been done and the curves published. The CERL Flue Dust Monitor fails with very fine dust whereas the optical density meter fails with very coarse dust. Physicists should read the original paper in order to appreciate the difficulties of trying to get a linear response for all sizes of particle. It can be mentioned as entertainment that in contrast to such carefully-devised instruments a polythene bowl is a perfectly good collector of grit and dust if one wishes to identify rather than measure. It has in fact been used by Warren Spring scientists when investigating specific complaints in order to identify the offender.

Such complaints rank highly in enquiries to the Warren Spring Laboratory, the housewife being very aware of the layer of dust or grit on curtains, window-sills, laundry and the like. The material is irritating to teeth and eyes but is not usually regarded as important in relation to health. A maximum of 0·45 g $m^{-3}$ is accepted as a suitable level of emission but older power stations rarely achieve this. Nor do factories that burn coal that is not pulverized. Modern power stations achieve 99·3% efficiency in the removal of dust and

grit; so monitoring of these is hardly necessary. (The common way of eradicating grit and dust from the emergent gases is to precipitate them electrostatically.)

Smoke is the most obvious and widespread pollutant, the cause of the old peasouper fog. It is therefore the one against which the Clean Air Acts are mainly directed. It has been shown that smoke, far from being entirely carbonaceous, can contain sulphuric acid, ammonium sulphate, traces of metals (up to 1 $\mu g\ m^{-3}$ has been measured) and so on. There is a correlation between the darkness of smoke and the cleanliness of buildings and textiles. There are fairly sound reasons for thinking that black material correlates with effects on health. Thus by chance and not because of knowledge the measurement and prevention of dark smoke has turned out to be the most important of anti-pollution activities. It is smoke that is measured by routine for the National Survey. Moreover it is smoke with which ordinary people are offenders, for domestic fires emit in Britain, even now in 1970, about seven times as much smoke as industrial sources.

Complete combustion would eliminate all the visible parts of smoke. Apart from this one obvious way to achieve amelioration is to invoke the principle of infinite dilution, i.e. perfect dispersal into the atmosphere at large. This can be done for industrial establishments by making chimneys very high and ensuring that the smoke is hot and is leaving the chimney at sufficient speed. The faster it emerges the less it is bent downwards by wind. Formulae connecting speed of emission, temperature, wind speed and shape of emergent plume were established years ago. The work on chimneys has been done at the NPL on scale models in a wind tunnel, for the meteorological conditions are important and they vary. In this way the correct height and positioning of the chimney and the speed of efflux are arrived at for the prevailing winds. A typical result of an investigation in 1964 (that for the development scheme for the Western General Hospital in Edinburgh) is: 'A 200-ft chimney operating with an efflux velocity of 40 feet per second in position G would provide complete freedom from pollution of the site and adjacent residential property for all but a total of 10 days per year when the direction of the wind is from 315° to 335°.'

Town-planners, however, do not like high chimneys; in any case they are expensive to build. The National Research Development Corporation has supported research by the British Hydromechanical Research Association to get the effect of height with a lower chimney. It is done by an ingenious fluidic switch—the smoke stream makes the switch—to pulse smoke out as smoke rings, which in still air rise to about twice the height of the chimney. The device is simple and novel in principle but involves ancillary equipment. A small one has been demonstrated at the BHRA laboratory at Cranfield.

High dispersal is obviously a good way of getting rid of smoke (and any gaseous pollutant) so that it does not pollute at ground level, but there is still the problem of existing industrial installations and domestic fires, which cannot have high chimneys. The solution here is to make sure that the smoke is not emitted at all, which means either a scheme for perfect combustion or the use of fuel that will not produce smoke. But this involves control and so smoke must be measured. Here it is extraordinary that an old system, modernized and modified, has persisted. The old Owens system drew polluted air through filter paper and assessed the blackness deposited by means of a series of ink blobs! Today the same technique is used but the darkening of the paper is measured by optical techniques. The transmitted light could be measured but this is not considered as consistent as the measurement of the light reflected. Obviously the time of exposure, the volume of air taken through, the size and quality of filter paper and the type of reflectometer are all important factors. Some automation of the routine is possible by the passage through the exposure cell of a continuous strip of paper held for a certain predetermined time in each position. The stain is a heterogeneous mixture of dust, products of combustion and other emissions. The variation from place to place and time to time makes it impossible to get a measurement of the smoke in $\mu g\ m^{-3}$ directly from the darkness of the stain. However, an international calibration curve has been obtained, by careful and tedious readings over a range of exposures, in Holland, France and Britain (Warren Spring). (The type of reflectometer and filter paper used makes variations of the calibration curve but readings of one type can

be converted into readings of another.) A 'standard smoke' was agreed which would make a concentration on the paper of 20 $\mu g\ cm^{-2}$ have a darkness index of 25 (100—reflection factor). With this it was possible to relate the calibration curves and normalize them at the chosen arbitrary figure. Thus it is possible to have an international measurement standard for smoke. Further experiments involved the measurement of the actual weight of deposited matter corresponding to an observed reflectivity. This nice piece of elementary physics makes comparable smoke concentrations anywhere The calibration technique devised at the Warren Spring Laboratory was accepted by OECD and has led to a British Standard Specification.

Sulphur dioxide is the prevailing pollutant that cannot easily be removed from combustion effluents (washing has already been mentioned but it lowers the temperature of the effluent and reduces its speed). Owing to its harmful effects on buildings and possibly on health, especially when sulphuric acid is produced, and to the possible creation of mists, it is considered an important pollutant and is measured for the National Survey. Several methods have been developed to get rid of it before emission of effluent. One was the Reinluft method tried, at first with optimism, at the Warren Spring Laboratory—it adsorbed the $SO_2$ on activated carbon—but eventually given up. Another method is the conversion to sulphuric acid. There is, for example, the Cat–Ox system developed and tried at an American power station by Monsanto Enviro-Chem Systems and made and installed in Britain by Simon–Carves Chemical Engineering. It is an elaborate system and obviously costly to install though with very sulphurous oil as a fuel the acid made might compensate for the cost. The general opinion of British air pollution scientists is that no system has yet proved itself to be effective and cheap and the line of progress of the future is to disperse $SO_2$ pollutant. For this purpose the Warren Spring Laboratory has several studies going on into the pattern of diffusion of pollution in cities. The standard method of measurement of $SO_2$ is a manual one. The polluted air, free of smoke, is drawn into hydrogen peroxide and the quantity of sulphuric acid produced is weighed. It is crude but simple.

A remarkable development at the Warren Spring Laboratory is the automation of measurements. The aim is eventually to be able to do automatic data logging and processing by computer. This means of course automatic sensings of pollutants and their conversion into signals. The method of sensing for $SO_2$ is to measure the change in conductivity of a hydrogen peroxide cell as sulphuric acid is formed. Plessey is concerned in one such automatic sensor at the Laboratory's Islington out-station. (It enables warnings of high pollution to be given publicly.) A project for the investigation of the spread of pollution from the runways at London Airport is involved in the development of automatic sensing of wind direction and speed as well as of $SO_2$ concentration. Shipton Automation is involved in the general research, being experts at answering-telephone technology. The $SO_2$ automatic sensor developed at the Warren Spring Laboratory is to be supported by NRDC and exploited commercially. The first stage of automation is the simple one of being able to interrogate an out-station by means of the ordinary telephone line. The more sophisticated system of the future will have out-stations with sensors for CO and other pollutants automatically relaying measurements to a central station. In connection with this there would be a warning system when certain levels of particular pollutants were measured. Dynatel and D Mac have also been so far involved in Dr J. Parker's work still continuing at the Warren Spring Laboratory. Such a network of communication and processing is not a new idea. Philips of Eindhoven has a system it has tried out for $SO_2$ near Rotterdam. Philadelphia is to have a computerized system costing 250,000 dollars. There is a small network scheme in Germany, another in Japan, others in the USA. So the system of automatic monitoring and early warning is very much a development for tomorrow. Dr J. Parker believes that the system being developed at the Laboratory will cost only a fraction of any of the others and be efficient and versatile.

The type of reaction used for the sensing and measuring of individual pollutants must be mentioned. CO is measured by infrared absorption. So is $CO_2$ by means of a different wavelength. The nitrogen oxides are measured in a two-stage dye reaction to distinguish

NO from $NO_2$. Methane is measured by flame ionization. Smoke and $SO_2$ have already been mentioned. What is needed is individual specificity, sensitivity and great reliability. It is the physicist on whom one will rely.

The UK Atomic Energy Authority's work is also concerned with instrumentation—its nephelometer for measuring atmospheric scattering and visual range was described in *Journal of Physics* E: *Scientific Instruments* 1970 Vol. 3. But whereas the scientists at the NPL and the Warren Spring Laboratory and CERL lean towards engineering and practical usable measurement, those at Harwell lean towards more fundamental physics. The properties of aerosols, the formation of mists, the photochemistry of $SO_2$; these are typical research projects. One of the Harwell projects was the investigation of mists under anti-cyclonic conditions on Tees-side, found to be related to ammonium sulphate. It is at Harwell that one finds the elaborate instruments. One more topic may be mentioned. It is that of the adsorption of $SO_2$ on particles. It is now agreed that $SO_2$ is not a major respiratory pollutant when experiments are done with air–$SO_2$ mixtures, for the gas is almost all removed in the upper respiratory tract. But if $SO_2$ is adsorbed on particulates that are in smoke, for example, these particles could penetrate to the lower lung and then release the $SO_2$. There might therefore be a synergistic effect accounting for the fact that some people do suffer respiratory trouble from pollution.

Nothing has been said about the work going on to deal with pollution caused by road vehicles. The subject deserves a separate article but has here been included in the general topic of air pollution. The measurement techniques are of course much the same, for the pollutants are the same except that the CO emission is higher than for industrial and domestic effluents and except for higher concentrations of nitrogen oxides and so on. All the road-vehicle pollutants come from incomplete combustion and intensive work is going on about this in America. One way of increasing the combustion is to recirculate gas from crankcase to cylinder head and at the same time convert CO in the exhaust to $CO_2$ by catalytic oxidation. These are not all very successful yet—the heat can melt the catalyst for example. Further experiments are concerned with the removal of aids to performance such as lead compounds. No doubt the research going on will produce some answers. The problems it seems will be those of cost and loss of power.

With all this work going on—and only a fraction has been mentioned—one is entitled to ask: What is the result? Are we getting anywhere? Figures given in table 2 have shown the fall in smoke emission since 1938. Dr S. R. Craxford and others at the Warren Spring Laboratory have produced data that also show significant trends. In carefully-derived data from the National Survey and from the statistics from the Ministry of Fuel and Power, he has shown that since 1951 the emission of smoke in the UK has been steadily falling, being 2·36m tonnes in 1951 and 0·93m tonnes in 1968. The White Paper of 1967 on Fuel Policy gives the 1975 figure as 0·60m tonnes. The yearly emissions (from the fuel consumed) have been plotted and compared with the measured concentrations and there is a remarkable correlation, the average concentration having fallen from about 170 $\mu g\ m^{-3}$ in 1958 to 70 $\mu g\ m^{-3}$ in 1968. A similar pair of graphs for London up to 1968 show that the concentration fell from 275 $\mu g\ m^{-3}$ in 1952 to 40 $\mu g\ m^{-3}$ in 1968, a fall of about 85·5%. There are, however, some remarkable differences in different regions, as already mentioned. The $SO_2$ concentration for the country as a whole has fallen very slightly over the years since 1963 as natural gas and nuclear energy have taken the place of coal and oil. It is further seen that industry as a whole makes 'only a very modest contribution' to ground-level concentrations of pollutants because of improved knowledge of how chimneys should be built and sited On other pollutants there are no comparable data over the years except for those from road vehicles, and these, as everyone knows, are increasing rapidly

The first conclusion from all this would seem to be that the Clean Air Acts have been singularly successful. The ironic truth is, however, that the decrease would probably have gone very much the same way without any legislation at all and what the Acts did was give a forceful push in the right direction. For changing social customs have made people give up coal and turn to gas and electricity for heating, or to smokeless fuels, and have made the wealthier

turn to central heating based on gas or oil. This social change will continue and the $SO_2$ pollution will fall as natural gas and nuclear-generated electricity are more used and high chimneys are used for dispersal. The other pollutants will get more attention. It is not possible to forecast so optimistically for pollution from road vehicles. People will not willingly give up their hardly-won motor-cars. So unless there is legislation and increasedly successful research the problem of pollution by cars will remain. It is a matter of urgency for physicists and motor engineers.

# Book Reviews

***Introduction to Coherent Optics and Holography***, by G. W. STROKE. (New York: Academic Press, 1969.) Pp. xvi+358. £5 7*s*.

As is well known, the principles of holography were first propounded in 1948 by D. Gabor, in connection with the technique of microscopy by wavefront reconstruction. However, the widespread applications of this method, and of the many other potentialities of coherent optics, had to await the development of the laser to provide light of adequate intensity, monochromaticity and coherence for their realization. Only in recent years, therefore, has holography been seen to have numerous important applications in science and technology.

The book under review was, in its first edition, generally recognized as being the first to be principally devoted to this subject. Although no longer alone, it remains an important standard work, compiled by an outstanding contributor in the field.

The first five chapters consist of an introduction, followed by a treatment of diffraction, image formation in non-coherent light, coherence characteristics of light and image formation in coherent light. The mathematical level of the theoretical treatments will be readily followed by most readers, and fortunately many of the errors which characterized the first edition have been removed. In some cases the method of revision leaves something to be desired: the theoretical treatment of the intensity interferometry experiments of Hanbury, Brown and Twiss, for example, which was incorrectly presented in the first edition, has simply been in large part omitted.

For many readers, chapter six, entitled "Theoretical and Experimental Foundations of Optical Holography (Wavefront-Reconstruction Imaging)", will be of major interest, containing as it does much useful material. The final chapter of the main text of the book is short and deals with Fourier transforms, convolutions, correlations, spectral analysis and the theory of distributions. It is a pity that the last topic has not been dealt with in more detail.

A major innovation in the second edition is the inclusion of an updating appendix by I. P. Nalimov, entitled "Applications of Holography", which though sketchy contains many important references. There are some curiosities, presumably arising in the translation from Russian: on pp. 211 and 212, for example, there is discussion of a laser *resonator* with a *non-reflecting* mirror (reviewer's italics)!

There is a substantial bibliography and a new photographic appendix following which Gabor's three pioneering papers of 1948, 1949 and 1951 are reprinted in full.

No doubt because of the need for rapid publication, the book presents a somewhat untidy appearance, detracting from its readability; the typesetting is often uneven, the range of type faces is ill-chosen and the figure captions are not adequately differentiated from the main text. Nevertheless, Professor Stroke's book is of considerable value, and is recommended to all who are interested in the broad fields of application now open for holography and related techniques.

C. J. MITCHELL

# Forthcoming courses and conferences

### NON-DESTRUCTIVE TESTING OF MATERIALS

**26–30 April 1971 at Brunel University**

This is a one week course. The topics covered include the origin and nature of defects, radiological methods, ultrasonic testing, magnetic particle techniques, penetrant methods, eddy currents and other electrical phenomena, and leak detection.

Details can be obtained after 1 January 1971 from Mr J. Blitz, Department of Physics, Brunel University.

### AN INTRODUCTION TO AERODYNAMICS

**23–27 November 1970 at Cranfield Institute of Technology**

A short course designed for non-specialists.

Details can be obtained from The Registrar, Cranfield Institute of Technology, Cranfield, Bedfordshire.

### INTERNATIONAL SCIENCE, RESEARCH AND TECHNOLOGY EXHIBITION AND CONFERENCE

**6–10 October 1970 at the National Hall, Olympia, London**

Conference sessions are being organized by IPPS on 7, 8 and 9 October. The purpose of these sessions will be to identify from the variouse technologies and disciplines such findings as may be of value beyond their research context.

# Sources of Reviews

**65–80** Direct current electroluminescence in ZnS by A. Vecht and N. J. Werring was published previously in *J. Phys.* D: *Appl. Phys.*, **3**, 105–20.

**81–95** The physics of Schottky barriers by E. H. Rhoderick was previously published in *J. Phys.* D: *Appl. Phys.*, **3**, 1153–67.

D. TABOR

Surface Physics, Cavendish Laboratory, Cambridge

# The hardness of solids

**Abstract.** This review is concerned with the basic physical meaning of hardness. It is shown that indentation hardness of ductile materials is essentially a measure of their plastic properties. There is a large hydrostatic component of stress around the indentation and since this plays no part in plastic flow the indentation pressure is appreciably higher than the uniaxial flow stress of the material. For many materials it is about three times as large but if the material shows appreciable elasticity the yielding of the elastic hinterland imposes less constraint on plastic flow and the factor of proportionality may be considerably less than 3.

Indentation hardness depends on the time of loading and on the temperature and the information so obtained may be correlated with the creep properties of the solid. At very high rates of loading as in dynamic measurements it is possible to calculate the dynamic hardness simply from the velocity of impact and rebound.

With brittle solids the high hydrostatic pressures around the deformed region are often sufficient to inhibit brittle fracture. Under these conditions both indentation and scratch hardness are essentially a measure of the plastic rather than the brittle properties of the solid. This provides a simple physical basis for the Mohs scratch hardness scale.

The review also deals with the relation between Brinell, Vickers and Rockwell hardness, the hardness anisotropy of single crystals, microhardness and the hardness of very small volumes, the indentation hardness of polymers and rubbers, and concludes with a brief discussion of the relation between hardness and molecular structure.

---

Hard is what is dense, and soft what is rare . . . Hard and soft as well as heavy and light are differentiated by the position and arrangement of the voids. Therefore iron is harder and lead heavier.

Democritus (*c.* 460–370 B.C.), quoted by Theophrastus, *de sensu* 61–62.

## 1. The concept of hardness

Broadly speaking our ideas concerning hardness are based on the resistance a solid shows to local deformation. With metals we press a hard indenter into the surface and measure the size of the permanent indentation formed; a similar type of test is used for polymers and other materials that flow plastically. With rubber-like materials we press an indenter into the material and assess how far it has sunk in under load. With minerals and brittle solids the classical method is to scratch one mineral by another of specified hardness number. All these tests are essentially means of giving a quantitative value to the material's ability to resist local deformation. In this article we shall show that all these hardness values are in fact a measure of certain bulk properties of the solids.

## 2. Hardness of metals

### 2.1. *The most common methods of measuring hardness*

The first of the modern techniques for measuring the hardness of metals was due to Brinell. The origin of the test may be of some interest to industrial scientists. Brinell was Technical Manager of the Fagersta Bruks steel works in Sweden at the turn of the century. He received complaints that a new batch of steel was not as 'good' as a previous batch, although the exact physical property being denigrated was not at all clear to the customer. Brinell thought of a very simple test to compare the satisfactory and unsatisfactory batches. He took a steel plate of each batch, placed a hard steel ball between them and squeezed them together in a vice. The comparative sizes of the dents showed whether one steel was 'harder' than the other.

In the standard Brinell test (Brinell 1900) a hard steel ball (usually 1 cm in diameter) is pressed normally on to the surface of the metal under examination. The load $W$ varies from 500 kg for soft metals to 3000 kg for hard steels and is usually applied for a standard period of 30 s. It is then removed and the diameter $d$ of the indentation measured. If there is negligible friction between ball and indentation it is easy to show that the true mean pressure over the surface of the indentation is

$$p = \frac{W}{\text{Projected area of indentation}} = \frac{4W}{\pi d^2}. \tag{1}$$

However, Brinell noticed that for most materials the pressure $p$ increased as he made the indentation larger. As we shall see this is due to work-hardening of the metal by the indentation process itself. In order to provide a hardness value which would show a much smaller variation with indentation size Brinell, therefore, chose as his definition of Brinell hardness number (BHN)

$$\text{BHN} = \frac{\text{Load}}{\text{Curved area of the indentation}}. \tag{2}$$

For indentations which are not too deep the difference between the BHN given by equation (2) and the true contact pressure (sometimes known as the Meyer hardness) given by equation (1) is not more than a few per cent. In the extreme case if the indenter is pressed into the metal up to its diameter the BHN will of course be only one half the Meyer hardness. In fact the general recommendation is that the chordal diameter $d$ of the indentation should be between 0·3 and 0·4 of the diameter $D$ of the indenter. Under these conditions

$$\text{BHN} \simeq 0{\cdot}97p. \tag{3}$$

It is seen that the Brinell hardness and the Meyer hardness $p$ have the dimensions of pressure. They are generally expressed in kilograms per square millimetre. Typical BHN values are 4 for lead, 6 for tin, 50 for annealed copper, 95 for brass, 150 for mild steel, 350 for tool steel. If as is customary the indenter used is of ball bearing steel (it may be shown by other tests that this has a hardness of about 900) it is advisable not to use it for testing metals of hardness greater than about 400 as otherwise some permanent deformation of the indenter itself may occur.

In the Vickers test (Smith and Sandland 1925) a square based diamond pyramid is used as the indenter. This can be used to determine the hardness of extremely hard materials. Again the Vickers Diamond Pyramid Hardness Number (DPH) is defined as

$$\text{DPH} = \frac{\text{Load}}{\text{Pyramidal area of the indentation}}, \tag{4}$$

$$= 1{\cdot}854 \frac{\text{Load}}{(\text{Length of diagonal})^2}. \tag{4a}$$

In the Knoop test (Knoop *et al.* 1939) a diamond indenter is used but its shape is such that it forms an elongated pyramidal impression. It is very useful for studying the effect of crystal orientation on hardness. The Knoop hardness number is again defined by equation (4). The Brinell, Vickers and Knoop indenters give hardness values which in most cases are close to one another.

The Rockwell test uses a spherical or a conical diamond indenter with a spherical tip. A small preliminary load is applied, then the main load is applied and removed and the depth of the residual indentation measured. This is shown on a dial gauge and is quoted as a Rockwell number. The test is very rapid and does not involve the optical measurement of the indentation diameter. For this reason it is a very convenient instrument in industry particularly if it is used mainly to check whether components satisfy a given specification. Depth measurements have, however, a number of basic weaknesses arising from 'piling-up'

and 'sinking-in' of material around the indentation and from the elastic recovery of the indentation when the load is removed (see later). A rapid test which avoids this defect could be a great asset to industry. One possible development along these lines which shows considerable promise is that due to Kleesatel (1966): the area of the indentation formed under load is deduced from the acoustic impedance of the indenter-metal interface. The hardness is again shown as a number on a dial but it is based on the area of the indentation, not on its 'recovered' depth.

Finally we may mention a method due to Bierbaum (1920) in which a fine diamond is dragged over the surface under a specified normal load and the width of the groove measured. This is essentially a research tool and provides a very useful technique for comparing the hardness of different portions of a surface. For example it can be used to study the hardness of different grains and of grain boundaries themselves.

A very full account of hardness equipment and mode of operation is to be found in the book by Weingraber (1952).

Most of the indentation methods used in industry operate at loads of the order of tens or hundreds of kilograms and the indentation size is of the order of a mm in diameter. The whole process may, of course, be miniaturized. If the indenters have very well-defined profiles, loads of the order of gms may be used. The indentations are then of the order of micrometres in diameter and high power microscopes must be used. Such tests are known as *micro-hardness* measurements and provide a valuable means of studying the deformation properties of very small regions. One of the major difficulties is that spurious values may be caused by vibration.

### 2.2. *The size of the specimen: friction between the indenter and the specimen*

Two technical points may be mentioned. The specimen must be appreciably larger than the indentation so that boundary effects do not influence the results. In general, if the diameter of the indentation is $d$ the distance of the indentation from the edge of the specimen (or from another indentation) should not be less than 3 to 4$d$. Similarly the thickness of the specimen should be at least three times the diameter (or eight to ten times the depth of the indentation) otherwise the indenter will 'sense' the presence of the surface on which the specimen rests. This is particularly important in measurements of the hardness of thin coatings.

The friction between the indenter and the specimen gives a higher value of the indentation hardness. Some theoretical calculations for two-dimensional wedges have been described by Grunsweig *et al.* (1954). More recently Johnson (1970) has re-examined the problem and shown that friction can appreciably change the *pattern* of deformation below the surface; however, it has relatively little effect on the indentation pressure itself. For example for a wedge of semi-angle 70° and a coefficient of friction of $\mu = 0{\cdot}1$ the indentation pressure is increased by about 10%. No solution exists for an axially symmetric indenter though one would expect the effect to be less marked. As a very crude approximation we may express the indentation pressure $p$ when the coefficient of friction is $\mu$ by a relation of the form

$$p = p_0(1 + \mu \cot \theta)$$

where $p_0$ is the value for frictionless indentation and $\theta$ the semi-apical angle of the indenter. For a Vickers diamond pyramid where the average value of $\theta$ is about 70°, this gives

$$p = p_0(1 + 0{\cdot}4\mu).$$

Consequently for a value of $\mu \simeq 0{\cdot}1$, a reasonable value for surfaces contaminated in the atmosphere, the observed indentation pressure is about 4% greater than the true value.

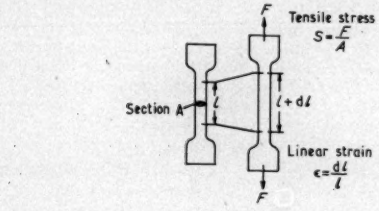
In a later section we shall see that the friction can be very important in hardness measurements at extremely small loads.

### 2.3. *The stress–strain characteristics of metals*

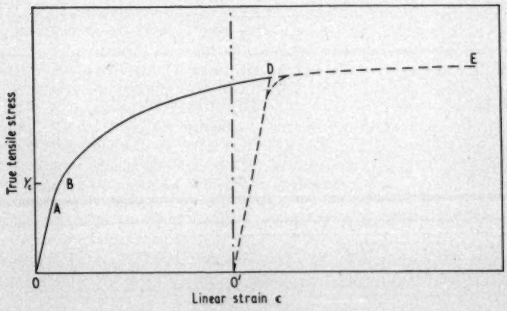
All the indentation hardness methods we have described are concerned with the permanent impressions formed in the metal. Although, as we shall see, some elastic effects may

be involved the overriding process is the plastic flow of metal around the indenter. This implies that the mean pressure over the indenter is connected with the plastic rather than the elastic properties of the metal. In what follows we shall in fact show that the hardness is directly related to the plastic stress–strain characteristics of the metal.

We consider the behaviour of a rod of metal subjected to uniaxial tension. We apply a given load and measure the extension of the rod. If the original length is $l$ and the extension is $dl$ the ratio $dl/l$ is known as the linear or engineering strain $\epsilon$. We express the tensile stress as the tensile force at any instant divided by the cross section of the specimen at that instant. This is the true stress $S$ as distinct from the engineering or nominal stress in which the force is divided by the initial cross section of the specimen. We need this quantity since it provides a true measure of the stress producing plastic flow.




(*a*)




(*b*)

Figure 1. (*a*) Tensile specimen, (*b*) tensile experiment in which the true stress $S$ (not the nominal stress) is plotted against the linear strain $\epsilon$.

The results obtained when $S$ is plotted against $\epsilon$ are shown in figure 1. Over the range OA the deformation is elastic. The strain is proportional to the stress and the deformation is reversible. If the material is ductile, elastic deformation will proceed until at some critical stress $Y_c$, known as the yield stress, the onset of permanent or plastic deformation occurs. If we continue along the plastic curve there is generally an increase in stress with deformation. In order to distinguish this from the initial yield stress $Y_c$, the stress is now generally referred to as the flow stress $Y$ and its increase is known as work-hardening.

Over a limited portion of the stress–strain curve the flow stress can be represented by a relation of the form

$$Y = b\epsilon^x \tag{5}$$

where $b$ is a constant for the metal and $x$ is a work-hardening index, usually ranging from a value 0 for metals which do not work-harden to about 0·6 for annealed metals which are capable of undergoing very marked work-hardening. A similar behaviour is observed if the metal specimen is compressed uniaxially between 'frictionless' anvils. It must be emphasized that equation (1) is a very crude relation, for it is clear that the plastic deformation curve does not start from the origin and $Y$ does not increase indefinitely with increasing strain; there is an asymptotic trend in $Y$ at high values of $\epsilon$. However, for our purposes the simple relation is adequate.

If an indentation experiment is carried out on a metal with a stress–strain curve similar to that shown in figure 1 the deformation associated with the plastic indentation process will increase $Y$ by varying amounts around the indentation. Before we can understand how to cope with this it is clearly desirable to consider the indentation behaviour of a metal which is not work-hardened by the indentation process itself. We may obtain such a material by starting with a metal such as that shown in figure 1 and deforming it heavily up to, say, the point D and then removing the applied stress. The specimen recovers elastically to the point 0′ where 00′ is the amount of permanent or plastic deformation. If this work-hardened specimen is now subjected to a renewed tensile test with the point 0′ as the new origin, the stress–strain curve follows the dotted curve 0′DE, i.e. the flow stress $Y$ along DE is almost constant. If an indentation is made in such a material the indentation process will have a negligible effect on $Y$. Consequently we may expect a relation between the indentation hardness and $Y$.

## 2.4. *Indentation of materials of constant yield stress*

A material obtained in the above way is a good approximation to the perfectly plastic material much favoured by theoreticians in studies of plastic deformation. Although in a perfectly plastic material the uniaxial stress to produce plastic yielding is a constant, $Y$, this only applies to the case of simple tension or compression in one direction. The stresses around an indentation are very complex. It is therefore necessary to know how an ideal plastic material will deform under combined stresses. One hint to the solution of this problem is provided by the observation that hydrostatic pressure itself plays no part in producing plastic flow. If, for example, a cylindrical specimen that would yield plastically under a uniaxial tensile (or compressive) stress $Y$ is subjected to a hydrostatic pressure of amount $Y$ or $2Y$ or $3Y$ no plastic flow occurs; a superposed uniaxial tensile (or compressive) stress $Y$ must still be applied if plastic flow is to take place.

Now analysis shows that the only part of a stress field that is unaffected by hydrostatic pressure is a shear stress. We conclude that plastic flow is associated with a critical resolved shear stress. This is fully supported by microscopic studies which show that plastic deformation is always accompanied by slip of atomic planes over one another. In its analytical form this leads to two slightly different criteria for plastic flow: one is a maximum shear–stress criterion (Tresca *et al.*), the other a criterion of maximum shear–strain energy (Huber *et al.*). Both agree that hydrostatic stress plays no role in plastic flow and provide a means of determining the behaviour of material under complex stress situations. The full analysis of the indentation process is far too involved to be given here but we may summarize the main conclusion descriptively by saying that nearly two-thirds of the mean pressure of contact is in the form of a hydrostatic pressure and plays no part in producing plastic flow; only one-third remains effective for this purpose (Hill 1950, Tabor 1951). Consequently if $p$ is the mean pressure between the indenter and the indentation $\frac{1}{3}p \simeq Y$ or

$$p \simeq 3Y. \tag{6}$$

This is the basic equation for all indentation measurements. It is necessary to emphasize at this stage the conditions under which it is valid. First the solid is assumed to be isotropic. This implies that, in practice, the metal must be polycrystalline without any preferred orientation. Secondly, it is assumed that it is fully work-hardened and has a constant yield stress $Y$, i.e. there is no difference between yield stress and flow stress. In practice this means that if the material work-hardens appreciably or shows an upper yield point, some other approach must be used. Thirdly it is assumed that elastic deformation is unimportant. This is generally true for most metals except very hard steels; it becomes increasingly inaccurate for materials for which $E/Y$ (where $E$ is Young's modulus) becomes less than about 100. In that case elastic yielding of the hinterland imposes less constraint on the plastic zone and $p < 3Y$. We shall discuss these issues in greater detail in later sections of this article.

### 2.5. *Vickers hardness of materials of constant yield stress*

The Vickers indenter is a square-based pyramid made of diamond in which the angle between the opposite faces is 136° (figure 2). Plasticity theory suggests that the mean yield pressure will depend somewhat on the angle of the indenter, but the relation between $p$ and $Y$ will be close to that given in the above equation. This conclusion may be examined by work-hardening metals as far as convenient, until a portion of the stress–strain curve is

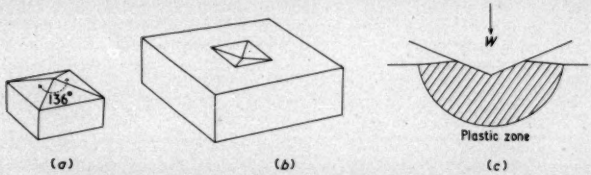



Figure 2. (*a*) Standard Vickers diamond pyramid indenter, (*b*) the indentation it produces, (*c*) the plastic flow around the indentation as deduced from theory.

reached where $Y$ is substantially constant. Vickers indentations may then be made and the mean pressure $p$ over the indentations calculated. This is given by

$$p = \frac{\text{Load}}{\text{Projected area of indentation}}.$$

Typical results are given in table 1; it will be seen that for a wide range of materials $p = 3{\cdot}2Y$. Since the Vickers hardness number DPH is defined as the ratio of the load to the pyramidal area of indentation (see equation (4)), DPH is less than $p$ by a numerical factor depending on the shape of the pyramid. For the standard Vickers indenter this factor is 0·9272. Consequently,

$$\text{DPH} = 0{\cdot}9272 \times 3{\cdot}2Y \simeq 3Y. \tag{7}$$

Hence the Vickers hardness number of an ideal plastic material is roughly three times its yield stress. Although many other factors are involved, it remains the basic relation between the indentation hardness and the bulk properties of a metal.

### 2.6. *Vickers hardness and ultimate tensile strength*

Since an ideal plastic material under tensile conditions passes its maximum nominal tensile stress $T_u$ as soon as plastic yielding begins, $Y$ is essentially the same as $T_u$. Hence to a close approximation $Y = T_u = 0{\cdot}33$ (DPH). In engineering circles the ultimate tensile

stress is usually expressed in tons per square inch, whilst the Vickers hardness number is expressed in kilograms per square millimetre. This reduces the conversion factor by 1·57; consequently

$$T_u \text{ (ton in}^{-2}) = 0{\cdot}21 \text{ (DPH kg mm}^{-2}). \tag{8}$$

This relation is widely used as a means of determining ultimate tensile strengths from hardness measurements for fully work-hardened materials. It should, however, be noted that this relation is not applicable to materials, such as cast iron, which are brittle in tension.

**Table 1. Relation between yield-stress $Y$ and indentation pressure $p$ (Vickers indenter)**

| Metal | $Y$ (kg mm$^{-2}$) | $p$ (kg mm$^{-2}$) | $p/Y$ |
|---|---|---|---|
| Tellurium–lead | 2·1 | 6·7 | 3·2 |
| Aluminium | 12·3 | 39·5 | 3·2 |
| Copper | 27 | 88 | 3·3 |
| Mild steel | 70 | 227 | 3·2 |

2.7. *Principle of geometric similarity*

The principle of geometric similarity states that if two indentations are made of the same geometric shape, then, whatever their size, the strain distribution and the stress distribution around the indentations will be geometrically similar. This is valid whether the material is fully work-hardened or undergoes work-hardening as a result of the indentation process, but it is only valid if the grain-size is so small that the material can be considered macroscopically as being uniform and homogeneous. In its simplest terms the principle implies that a large indentation is essentially a magnified picture of a small indentation, the strains and hence the stresses being the same at any geometrically similar region. Consequently the indentation yield pressure $p$, which is the mean pressure acting on the indenter, will be the same whatever the size of the indentation. It follows that for a pyramidal (or conical) indenter the hardness is independent of the size of the indentation. This is well borne out in Vickers hardness measurements where the hardness is generally found to be independent of the load.

We consider here two applications of this principle: the Vickers hardness of materials which work-harden and the Rockwell hardness test.

2.8. *Vickers hardness of materials which work-harden*

With a material which can work-harden, the indentation process itself will produce work-hardening, that is, it will produce an increase in the flow stress $Y$. With the Vickers pyramidal indenter the plastic strains produced will vary from point to point so that $Y$ will also vary from point to point. A detailed analytical solution of the problem is not possible but we may provide a very simple analysis in the following way. We assume that there is a representative value of the flow stress $Y_r$, which is related to the observed Vickers hardness by the relation $\text{DPH} = cY_r$, where $c$ again has the value of about 3. By an empirical method it is found that the indentation produces an average or representative strain $\epsilon_r$ (corresponding to $Y_r$) equivalent to an 8% tensile strain. Hence, if on the known tensile stress–strain curve of the metal we determine the flow stress for an additional 8% strain, the Vickers hardness value will be three times this value. Because of the principle of geometric similarity, this additional strain will be the same whatever the size of the indentation. It also turns out that the additional strain is roughly constant whatever the initial state of work-hardening of the specimen. This is shown in table 2 for annealed copper and steel specimens which were work-hardened by various strains $\epsilon_0$ and had their Vickers hardness measured. The flow stress at a strain of $(\epsilon_0 + 8)\%$ was determined from their stress–strain curves and compared with the observed Vickers hardness values. The agreement between the last two columns is fairly good. It should, of course, be borne in

mind that the addition of linear strains is not strictly valid. Indeed a more detailed analysis of these data by Voce (1951), using natural or logarithmic strains which are additive and a more sophisticated type of relation for the stress–strain curve, gives even better agreement than that recorded in table 2.

**Table 2. Relation between flow-stress $Y$ and Vickers hardness number, for metals which are work-hardened by the indentation process**

| Metal | Initial deformation $\epsilon_0$ (%) | $(\epsilon_0+8)$ (%) | $Y$ (kg mm$^{-2}$) at strain of $(\epsilon_0+8)$% | $cY$ (kg mm$^{-2}$) | Observed Vickers hardness number (kg mm$^{-2}$) |
|---|---|---|---|---|---|
| Mild steel | 0 | 8 | 55 | 159 | 156 |
| | 6 | 14 | 62 | 176 | 177 |
| | 10 | 18 | 66 | 190 | 187 |
| | 13 | 21 | 67 | 194 | 193 |
| | 25 | 33 | 73 | 211 | 209 |
| | | | | $c=3{\cdot}0$ | |
| Annealed copper | 0 | 8 | 15 | 45 | 39 |
| | 6 | 14 | 20 | 60 | 58 |
| | 12·5 | 20·5 | 23·3 | 70 | 69 |
| | 17·5 | 25·5 | 25 | 75 | 76 |
| | 25 | 33 | 26·6 | 80 | 81 |

More recently Dugdale (1958) has analysed the relation between hardness and flow stresses in much greater detail. His results show that in order to obtain consistent agreement a very much more complicated treatment is necessary. However, our greatly simplified analysis shows that hardness values obtained with the Vickers indenter provide a measure of the plastic flow stress of the metal as augmented by the indentation process itself.

2.9. *The Rockwell test using a conical indenter*

The indenter is first located on the specimen with a 'minor' load of 10 kg. The depth measuring dial indicator is set at zero. The 'major' load (usually 100 or 150 kg) is then applied for a specified short period and then removed. The indenter, under its minor load, rests at the bottom of the indentation and the dial indicator now records the Rockwell hardness number.

The basic connection between the Rockwell hardness number and the indentation hardness of the material is easily derived if we make a few simplifying assumptions. We ignore the presence of the spherical tip and assume that the indenter is a perfect cone. We ignore the effect of the 'minor' load and assume that the depth is essentially the same as if only the major load had been used. Finally we ignore pile-up or sinking-in of material around the indentation (see below) as well as shallowing of the indentation due to elastic recovery. Elastic shallowing can be particularly marked with hard steels (see later).

For a conical indenter the principle of geometric similarity applies so that the mean indentation pressure $p$ for a given material is independent of the load. If $d$ is the diameter of the indentation and a constant load is used in testing various metals, $p=k_1/d^2$. The depth of penetration $t$ is $\frac{1}{2}d\cot\theta$, where $\theta$ is the semi-angle of the cone. Consequently $t=k_2p^{-1/2}$ for a given cone. This means that the depth is smaller for larger values of $p$, that is, for harder materials. In order to provide a scale in which large-scale readings are associated with harder materials, the depth-measuring dial-gauge is run in reverse, that is, the depth of penetration is subtracted from a fixed zero value, say $C$. The observed Rockwell depth $R_C$ is then

$$R_C=C-t=C-k_2p^{-1/2}.$$

The apical angle of the cone is not widely different from that of the Vickers pyramidal indenter and the indentation pressure is roughly the same in both cases. We may therefore

put $p$ equal to the Vickers hardness DPH and so obtain a relation between the Rockwell $R_C$ scale and the Vickers hardness values.

$$R_C = C - k_2\,(\mathrm{DPH})^{-1/2}. \tag{9}$$

A typical calibration curve is given in figure 3. In view of the simplifying assumptions made the agreement is surprisingly good.

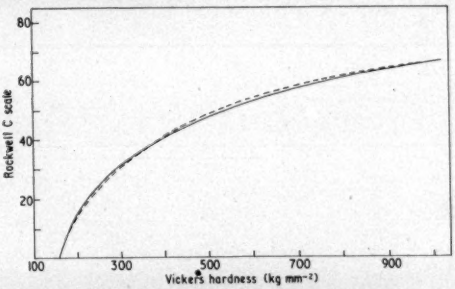



Figure 3. Comparison of Rockwell C hardness values and Vickers hardness. —— Experimental results, — — — — curve calculated according to the theoretical relation $R_c = 110\,(1 - 12{\cdot}4\,(\mathrm{DPH})^{-1/2})$.

2.10. *Brinell hardness measurements of fully work-hardened metals*

We consider first the indentation of a fully work-hardened metal specimen (i.e. of constant yield stress $Y$) by a hard sphere. At small loads the surface is deformed elastically the maximum shear stress occurring at a region within the metal a little way below the centre of the contact region. At some critical load this exceeds the critical resolved shear stress of the metal and a small amount of plastic flow occurs within the larger elastic hinterland (figure 4(*a*)). The first permanent indentation is very small and the contact pressure is of the order $p \simeq 1{\cdot}1\,Y$. As the load is increased the indentation grows in size, the plastic zone grows and the contact pressure increases. By the time the plastic zone has reached the free surface and has completely embraced the region around the indenter the contact pressure has reached a value of about $p \simeq 3\,Y$ (figure 4(*b*)). The behaviour now resembles that of a conical or pyramidal indenter so that the hardness is nearly the same as in a

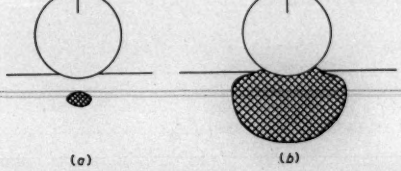
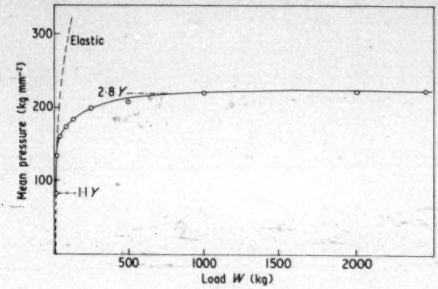
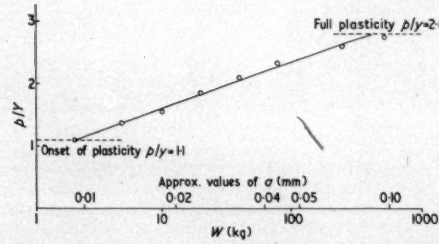



Figure 4. Deformation around a spherical indenter (*a*) onset of plastic flow below the surface, (*b*) at a higher load 'full plasticity' is reached and flow surrounds the indenter and extends to the free surface.




(a)




(b)

Figure 5. (*a*) Indentation pressure $p$ as a function of load $W$ for a hard steel ball of diameter 10 mm pressed into the surface of a fully work-hardened mild-steel specimen (yield stress $Y = 77$ kg mm$^{-2}$). (*b*) The same results plotted as $p/Y$ against ln $W$. This figure also shows that the radius $a$ of the indentation increases by a factor of about 10 as the deformation passes from the onset of plasticity to full plasticity.

Vickers test. In fact at this stage the hardness appears to be nearer $2{\cdot}8\,Y$ than $3\,Y$ and this seems to be generally true for spherical indenters. As figure 5 shows, the condition of full plasticity is reached at a load about 300 times that at which the onset of plastic deformation occurs. This corresponds to an increase in the diameter of the indentation by a factor of about ten. The hardness remains constant at higher loads.

The difference between the onset of plastic indentation and full plasticity has a direct bearing on the deformation of the indenter itself. Suppose the sphere has a yield or flow stress $Y_B$ and the metal a yield or flow stress $Y_M$. As indentation of the metal proceeds the contact pressure increases until it reaches a value $p_M$ of about $2{\cdot}8\,Y_M$. If at this stage the pressure exceeds $1{\cdot}1\,Y_B$, the ball indenter itself will begin to deform plastically. Consequently the ball will be permanently flattened if

$$p_M = 2{\cdot}8\,Y_M > 1{\cdot}1\,Y_B$$

i.e. if

$$Y_M > \frac{1}{2{\cdot}5}\,Y_B$$

or if

$$p_M > \frac{1}{2{\cdot}5} p_B$$

where $p_B$ is the yield pressure or hardness of the ball in the condition of full plasticity. This means that if the ball has a Vickers hardness of about 900 kg mm$^{-2}$ (a reasonable value for ball bearing steel) it should not be used to indent metals of hardness greater than about 400 kg mm$^{-2}$.

### 2.11. *Brinell hardness of materials which work-harden*

In a later section we shall consider the growth of the plastic zone in greater detail. In what follows we shall assume that we are always operating in the régime of full plasticity. If we indent a metal which work-hardens using a spherical indenter we find that the hardness increases with load. Unlike the pyramidal indenter used in the Vickers test the hardness is not independent of the size of the indentation. This is because indentations of various sizes formed with a given spherical indenter are not geometrically similar. In contrast to the pyramidal or conical indentations, a large indentation produced by a given sphere involves greater plastic strains than a small one, a corresponding increase in the effective flow stress, and hence an appreciable increase in the observed hardness. Thus the hardness is larger the bigger the indentation. Although this means that the Brinell hardness is *not* a single-valued quantity, its variation with load provides additional information about the properties of the metal.

Let us assume that the mean pressure $p$ over the indentation is related to some representative value of the flow stress $Y$, by a relation of the type $p = cY_r$, where $c$ again has a value of about 3. If $d$ is the chordal diameter of the indentation and $D$ the diameter of the indenter, the shape of the indentation is determined by the dimensionless parameters $d/D$. Thus the strains must be determined by this parameter. These will vary from point to point but we may again assume that there exists a representative strain $\epsilon_r$ corresponding to $Y_r$. A detailed study shows that to a first approximation $\epsilon_r$ is a linear function of $d/D$. As a first approximation it may be written as

$$\epsilon_r = 20\,\frac{d}{D}\,(\%). \tag{10}$$

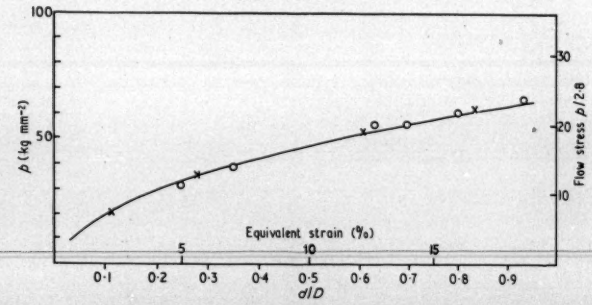



Figure 6. Indentation hardness $p$ of annealed copper as function of size parameter of indentation ($d/D$). The experimental hardness values are shown as circles and crosses. This may be converted into the stress–strain curve by putting the effective strain equal to $20(d/D)\%$ and the flow stress equal to $p/2{\cdot}8$. The full line is the true stress–strain curve obtained from 'frictionless' compression experiments.

The indentation pressure will be a numerical factor (of the order of 3) times the flow stress $Y_r$ corresponding to the indentation giving a representative plastic strain $\epsilon_r$. If, therefore, we carry out a series of indentations at increasing loads and plot $p$ against $d/D$ we may reconstruct the true stress–strain curve of the metal. We convert the $d/D$ values into the effective strain using equation (10) and the contact pressure $p$ into flow stress by dividing by 2·8. A typical result is shown in figure 6. This is a rough but practical way of determining (over the first 20% of strain) the stress–strain curve of small specimens which cannot be machined into tensile test-pieces.

This may be taken one stage further. If we assume that the stress–strain curve is given by equation (5) we may write

$$Y_r = b\left(\frac{kd}{D}\right)^x.$$

Hence

$$p = p_0\left(\frac{d}{D}\right)^x \tag{11}$$

where $p$ is the mean contact pressure and $p_0$ a suitable constant. Since, for an applied load $W$, $p = 4W/\pi d^2$, it follows from equation (11) that

$$W = \frac{Ad^{2+x}}{D^x} \tag{12}$$

or in the form that this was first derived experimentally by Meyer

$$W = \frac{Ad^n}{D^{n-2}} \tag{13}$$

where $A$ is a constant and $n$ is a constant known as the Meyer index (Meyer 1908). It follows, as Meyer showed, that if the load $W$ is plotted against the diameter $d$ of the indentation on log–log coordinates we should obtain a straight line of slope $n$. This is in fact so over a very wide range of experimental conditions and materials. The Meyer analysis yields values of $n$ which lie between 2 for fully work-hardened metals and about 2·6 for fully annealed metals. The above treatment shows why this is so and connects it with the work-hardening index of the metal. Of course this analysis assumes the validity of the power-law relation between stress and strain. It is therefore of limited applicability; nevertheless it provides a reasonably reliable means of assessing the work-hardening characteristics of a metal.

### 2.12. *The hardness of metals showing an upper yield point*

With metals which show a marked upper yield stress the hardness will depend on the flow properties of the material after it has been subjected to several per cent strain. Consequently the hardness will not correlate with the *upper* yield stress but with the lower yield stress (flow stress) at some appropriate value of strain (Dugdale 1958).

### 2.13. *Piling-up and sinking-in*

The simple theory of plastic indentation leads to two general predictions. The first is the relation between the indentation pressure and the flow stress: this has been discussed fully in the preceding sections. The second is the pattern of plastic flow around the indenter. For an ideal rigid-plastic solid, where elastic deformations are negligible and the yield stress is constant, the displaced material flows up around the indenter and forms a raised portion close to the indenter surface. This behaviour is, in fact, observed in the indentation of work-hardened metals: there is a 'pile-up' of metal around the edge of the indenter. With a spherical indenter the pile-up is symmetrical (see figure 7(*a*)). With a pyramidal indenter the metal is freer to move up the centre of the pyramidal faces and is more constrained at the edges: as a result it moves up further at the centre of the faces and so produces

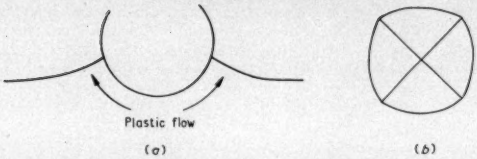



Figure 7. Deformation and flow of metal around an indenter for a work-hardened specimen. (*a*) For a spherical indenter there is a raised ridge or pile-up of material near the edge of the indentation, (*b*) for a pyramidal indenter the upward flow is easier along the centre of the pyramidal faces: this gives a barrel-shaped indentation.

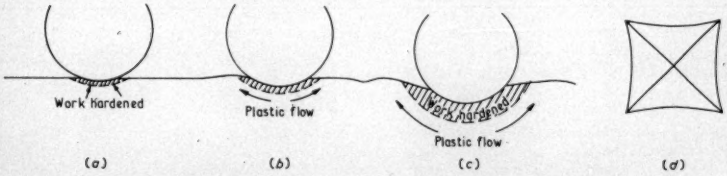



Figure 8. Deformation and flow of metal around an indenter for an annealed specimen. (*a*), (*b*) and (*c*) for a spherical indenter sinking-in occurs near the edge of the indentation, (*d*) for a pyramidal indentation the impression acquires a pincushion shape.

a barrel-shaped, instead of a square, indentation (see figure 7(*b*)). Clearly a small correction must be applied in calculating the Vickers hardness from the length of the diagonal.

With highly annealed metals the behaviour is strikingly different. When the indenter first begins to sink into the metal the material adjacent to the indenter becomes markedly work-hardened relative to the undeformed metal farther away (figure 8(*a*)). As a result, when the indenter sinks in farther, it carries with it the surrounding metal which acts as a sort of enlarged indenter and deforms the metal adjacent to it (figure 8(*b*)). The result is

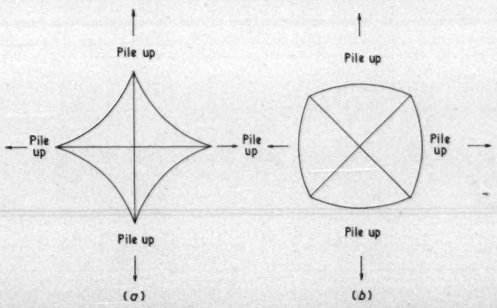



Figure 9. Schematic diagram showing indentation of rock salt with a pyramidal indenter; arrows represent directions of easy flow. This leads to a pincushion-shaped indentation in (*a*) and a barrel-shaped indentation in (*b*). Piling-up occurs in the same region in both cases.

that the displaced metal always appears to be moving out farther and farther away from the indenter itself. This leads to a depression of the metal immediately adjacent to the indenter and a slight piling-up some distance away (figure 8(*c*)). The same effect occurs with the Vickers indenter, but again the sinking-in is more marked at the centre of the pyramid faces; this leads to the characteristic pincushion type of indentation (figure 8(*d*)). This type of indentation is commonly observed with polymers. Here again a correction must be applied in calculating the Vickers hardness from the length of the diagonal.

With single crystals the behaviour is greatly complicated by the fact that plastic flow occurs only on a limited number of slip planes and that the critical resolved shear stress may differ greatly from one slip system to another, that is, there is a marked anisotropy in plastic properties. A typical example obtained on the (100) face of rock salt using a pyramidal indenter is shown schematically in figure 9. A pincushion impression is formed for one orientation of the indenter, a barrel-shaped impression if the pyramidal edges are rotated through 45°. The rock salt flows more easily along the directions marked with arrows and produces piling up in this direction whatever the orientation of the indenter. Similar results are generally observed with crystals of metals (Bückle 1959, Dyer 1961, Courtel 1961), and minerals (Bergsmann 1944). As we shall see below anisotropy of flow is of general occurrence in single crystals.

## 3. Size effects in the hardness of metals: anisotropy of single crystals

In this section we shall discuss three main size effects: first the influence of grain size on hardness; secondly the behaviour of single crystals and their anisotropic hardness properties; thirdly the effect of reducing the size of the indentation by working at very small loads, i.e. microhardness measurements.

### 3.1. *Grain size*

In the course of a study of the stress–strain properties of mild steels of various grain sizes Hall (1951) observed that the critical shear stress $\sigma$ in the lower-yield régime depended on grain size in a relatively simple way. If $l$ is the diameter of the grain

$$\sigma = \sigma_0 + \alpha_1 l^{-1/2} \tag{14}$$

when $\sigma_0$ is the yield stress of a large single crystal.† This increase is due to the resistance exerted by the grain boundary to the passage of slip from one grain to the next. It may be explained in terms of the stresses exerted by a pile-up of dislocations at the boundary. In a more recent paper Jindall and Armstrong (1967) confirmed this result and in addition showed that the factors influencing the magnitude of the effect are the number of slip systems, the shear modulus and dislocation locking within the individual grains. With 70 : 30 brass $\alpha_1$ remains constant with strain, but with pure metals $\alpha_1$ tends to decrease with increasing strain.

If hardness is a measure of the yield stress we should expect (as suggested by Hall (1954)) that the hardness would follow a similar law

$$p = p_0 + \alpha_2 l^{-1/2}. \tag{15}$$

Some typical indentation hardness results by Bassett and Davis (1919) and Babyak and Rhines (1960) where the indentation covered a large number of grains are reproduced in figure 10; they show that this relation is well obeyed. Jindall and Armstrong (1969) have studied this in greater detail and confirm the general validity of equation (15). However, there is not an exact correlation between equation (14) for $\sigma$ and equation (15) for $p$. This is not surprising since the strains produced by the indentation process will have a disparate influence on $\sigma_0$ and $\alpha_1$.

It appears that the increase in hardness with decreasing grain size is a genuine effect which may be described by a relation of the type given above. However, some materials such as

† Petch (1953) observed a similar dependence of fracture strength on crystal size.

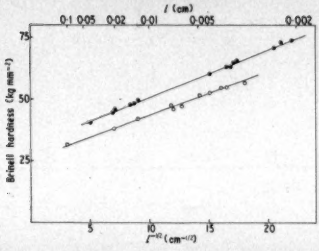



Figure 10. Brinell hardness plotted against $l^{-1/2}$ for polycrystalline brass where $l$ is the mean diameter of the crystallites. Results ● Bassett and Davis (1919), ○ Babyak and Rhines (1960).

α-uranium behave in a far more complicated way. Taplin (1966) suggests that this is because this material twins very readily under deformation and the twin composition planes themselves provide barriers of a type similar to those provided by grain boundaries.

### 3.2. *Hardness of single crystals: anisotropic behaviour*

As we saw in earlier sections, a fully work-hardened polycrystalline specimen of a pure metal behaves like a classical isotropic plastic solid: its hardness is approximately three times the uniaxial yield stress. A single crystal of such a material has an extremely low yield stress, yet the indentation hardness is not markedly less than that of the fully worked material. For example with aluminium the indentation hardness is of order 20 kg mm$^{-2}$ (yield stress about 1 kg mm$^{-2}$) whereas for a work-hardened polycrystalline specimen it is about 50 kg mm$^{-2}$ (yield stress about 16 kg mm$^{-2}$). Indeed Westbrook (1958) has suggested that for single crystals of the rock salt structure

$$p \simeq 30\,Y. \tag{16}$$

This may be due partly to a very rapid work-hardening rate in the single crystal. An additional factor appears to be the complex nature of the deformation around an indenter and the limited number of slip planes along which plastic flow can occur. The material is forced to deform in certain directions by the shape of the indenter and this reveals the constraints imposed by the slip system on the yielding of the crystal.

Anisotropy in hardness is very conveniently studied using a Knoop diamond indenter since it produces indentations which are seven times as long as they are wide (see figure 11). Earlier work by Daniels and Dunn (1949), Partridge and Roberts (1963) and Garfinkle and Garlick (1968) showed very clearly that anisotropy in hardness is determined by crystal structure and the primary slip systems. Figure 12 for example shows similar results more recently obtained by Brookes *et al.* on the (001) plane of MgO. The direction of the long axis of the indenter was first oriented parallel to a [100] direction and the hardness measured. It was then rotated through various angles and the hardness (on fresh portions of the crystal surface) determined. The results show that in the [110] direction the hardness is very much greater than in the [100] direction. Similar results (though not so marked) are observed on LiF which has the same structure and slip system. However, on the (100) face of a single crystal of $CaF_2$ (fluorspar structure) the maximum hardness is in the [100] direction, the minimum in the [110] direction the values differing by about 20%.

The anisotropy must clearly depend on the way in which the various slip systems in the bulk of the crystals respond to the applied load. If a tensile load $F$ acts along the axis of a cylindrical monocrystalline specimen of area $A$, rotation of the slip planes will generally

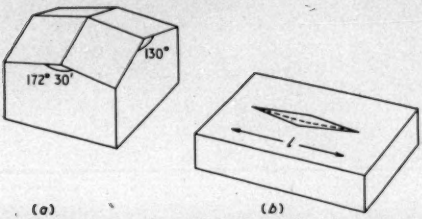



Figure 11. (*a*) Knoop indenter, (*b*) type of indentation produced. The indentation is in the form of an elongated pyramidal impression, its length being seven times its width. The Knoop indenter is particularly suited for studying anisotropy in hardness. It also generally gives more satisfactory indentations in brittle solids than the Vickers indenter.

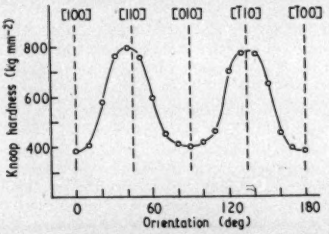



Figure 12. Knoop hardness values obtained on the (001) face of a single crystal of MgO, as a function of the orientation of the long diagonal of the indenter. The hardness in the ⟨110⟩ direction is far larger than in the ⟨100⟩ direction.

occur. If $\lambda$ is the angle between the stress axis and the slip direction and $\phi$ the angle between the stress axis and the normal to the slip plane, the critical resolved shear stress producing slip on the plane is given by the well-known Schmid and Boas relation

$$\tau = \frac{F}{A} \cos \lambda \cos \phi. \tag{17}$$

Daniels and Dunn (1949) suggested that the main factor involved in indentation was the stretching of the crystal as the indenter penetrated the surface. They therefore considered the crystal to be made of a whole series of cylinders lying parallel to the steepest slope of the indenter faces and subjected to tension. The constraint opposing rotation during indentation depends on the angle $\chi$ between the face of an adjacent indenter-face and the axis of rotation of the slip system. They suggest that the effective resolved stress capable of producing plastic yielding is given by

$$\tau_e = \frac{F}{A} \cos \lambda \cos \phi \cos \chi. \tag{18}$$

On this view the higher the value of $\tau_e$ the more readily slip will occur, i.e. the softer the material will appear to be. This relation explains some but not all of their data. In a later paper Feng and Elbaum (1958) and Garfinkle and Garlick (1968) assumed that the effective deformation stress was a compressive stress normal to the indenter faces. This again could not explain all the observed phenomena for crystals of different structures.

Recently Brookes *et al.* have re-examined the analysis of Daniels and Dunn and have shown that the effective resolved stress is determined not only by $\chi$ but by another angle $\gamma$ which represents the angle between the slip direction and a horizontal axis lying parallel to the indenter-face. They deduce that the effective resolved shear stress is

$$\tau_e = \frac{F}{A} \cos\lambda \cos\phi \tfrac{1}{2}(\cos\chi + \sin\gamma). \tag{19}$$

All these angles are determined by the geometry of the indenter and the slip systems in the crystals. Figure 13 shows the calculated values of $\tau_e$ obtained for the (100) face of a crystal with a rock salt structure where the slip system is $\{100\}\langle 1\bar{1}0\rangle$. In this figure $\tau_e$ is the

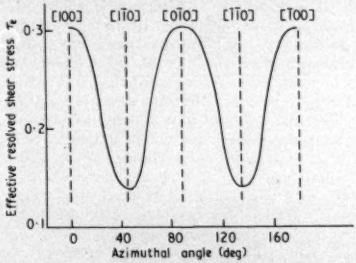



Figure 13. Calculated value of the effective resolved shear stress $\tau_e$ around a Knoop indenter pressed into a crystal with the rock salt structure. The greatest value is in the $\langle 100\rangle$ direction so that this should be the softest direction (compare with figure 12).

average for all the slip planes in the crystal. It is seen that the greatest value is in the [100] direction so that this would be the softest direction; the lowest value is in the [110] direction so that this would be the hard direction. The results agree well with figure 11 though it should be remarked that the absolute values are not always in close agreement. For example the hard directions in MgO are twice as hard as the soft; in LiF only 20% as hard. Nevertheless the broad pattern is well substantiated. Similar agreement is obtained for the fluorspar structure and for hcp crystals. In the latter case the authors show that for those materials for which the slip system is $\{1\bar{1}00\}\langle 11\bar{2}0\rangle$ (e.g. Ti, Zr) the hardest direction on the $(1\bar{1}00)$ face is the $\langle 11\bar{2}0\rangle$, whereas if the slip system is $\{0001\}\langle 11\bar{2}0\rangle$ (e.g. Mg, Zn, Co) this is the softest direction, the hardest direction being the $\langle 0001\rangle$.

The analysis of Brookes *et al.* is based on a number of assumptions concerning the factors which produce slip. They emphasize the tension on subsurface material whereas most indentation studies stress the role of compressive forces normal to the indenter faces. It is true that the maximum resolved shear stress for a tensile axis parallel to the steeper slope of the indenter face varies in almost the same way as the compressive stress normal to the indenter face: but the agreement is not exact. Brookes *et al.* have found that the observed anisotropy appears to agree better with the calculated tensile rather than the compressive stress and since this is at variance with current views it merits further examination. In addition, they use a direct averaging to find $\tau_e$, whereas one might expect the unfavourable orientations to be more heavily weighted. Nevertheless the general agreement with observation is very satisfactory and emphasizes the relation of the indenter-geometry to the slip-systems in the crystal.

### 3.3. *Microhardness: indentation hardness at very small loads*

Microhardness measurements are usually carried out with loads below about 50 g. In order to avoid the problems of 'full plasticity', it is usual to use a Vickers pyramid

indenter with a very carefully profiled tip. This has two further advantages: first it gives far more clearly defined boundaries than a spherical indenter; secondly it gives geometrically similar indentations so that if the material is of uniform hardness the hardness should be constant however small the load. In the load range used the indentation diameter is of order 1–10 micrometres. In some cases it is found that the hardness diminishes at very small loads but this is probably due to vibration which becomes increasingly important as the load is reduced. More generally there is an increase in hardness at small loads. This may be due to the method of preparing the specimen since the surface layers may be hardened by the polishing process (Bowden and Tabor 1964). Again some workers have suggested that there is elastic recovery of the indentation *diameter* when the load is removed but on grounds of geometric similarity this should not depend on the size of the indentation. Careful experiments described by Bückle (1959, 1960) and by Mott (1957) suggest that an increase in hardness at small loads is genuine and they attribute this, in somewhat different terms, to the limited number of dislocations in the small volumes being deformed.

If indentations could be made in a volume so small that it contained no dislocations we should expect the hardness to approach the theoretical value of a perfect crystal. If the shear modulus is $\mu$ this should correspond to a critical shear stress of the order $\frac{1}{30}\mu$ whereas for a pure crystal containing dislocations it is 10 or 100 times smaller. This would imply that a dislocation-free crystal should have a hardness 10 to 100 times greater than one containing dislocations. However, as pointed out above, the hardness of single crystals is greatly affected by the constraints imposed by the limited number of slip planes. All that one can say is that a very large difference in hardness should be observed. One might in fact expect the perfect crystal to have a hardness comparable with that of a fully work-hardened specimen where dislocation entanglement is so marked that the yield properties resemble those of a material which contains no dislocations.

A study along these lines has recently been carried out by Gane (1970). In most of his experiments, in order to avoid the complications due to oxide films, he used single crystals of gold. They were electropolished and then annealed. The indenter consisted of a fine conical tip of tungsten such as those used in field-ion microscopy and had a tip radius of the order of 1000 Å. The indenter was supported on a delicate galvanometer movement so that, simply by varying the current, loads ranging from 1 g down to 1 mg could be applied. These loads would be expected to give indentations 10 to 100 times smaller than those obtained in conventional microhardness measurements. The whole system was mounted in a scanning electron microscope so that the indentation could be viewed as it occurred.

The results fall into two classes. In one it was found that below a certain load no indentation or noticeable surface distortion occurred: above this load the indenter suddenly penetrated the gold surface. The local yield pressure $p$ when indentation commenced was of order 450 kg mm$^{-2}$ (see figure 14). Beyond this point the hardness fell rapidly with increasing penetration. At this stage the indentation had a diameter of about 1 micrometre. From the geometry of the indentation the representative strain could be deduced and the results are plotted in figure 15. It is seen that for a strain of about 20% the hardness is of the order of 80 kg mm$^{-2}$ which is perhaps twice the macroscopic hardness of a gold single crystal. The behaviour resembles that of a material possessing an upper yield stress followed by a relatively small lower yield stress. If the hardness is assumed to be about $5\tau$ where $\tau$ is the critical resolved shear stress, this gives for the maximum hardness a value of $\tau \simeq 90$ kg mm$^{-2}$ which is about $\frac{1}{30}\mu$.

The second class of behaviour was one in which no upper yield was observed. The hardness was almost constant at about 80 kg mm$^{-2}$ (broken curve in figure 15).

Similar results were obtained if a fine gold tip was pressed on to a hard flat anvil. In some cases an upper yield was observed ($p \simeq 450$ kg mm$^{-2}$) followed by a rapid drop; in other cases a low hardness value was observed even at the smallest loads.

Detailed studies show that the two classes depend on the presence or absence of a thin surface film formed by polymerization of the vapours present in the microscope. If the film is present the high hardness is observed: if absent no upper yield occurs. This is not

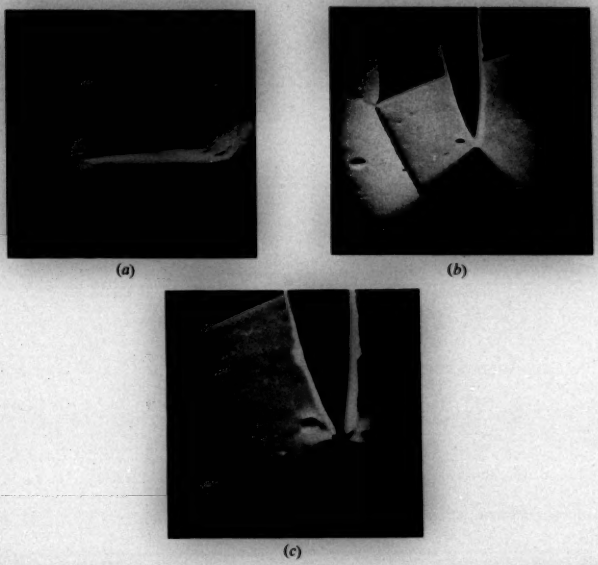


Figure 14. Scanning microscope studies of microindentations in a single crystal of gold using a fine tungsten tip as indenter. (*a*) Indenter resting in an indentation already formed with a Vickers pyramidal diamond at a load of 10 g. (*b*) Indenter pressing on surface at a load just insufficient to cause indentation. Contact pressure is about 450 kg mm$^{-2}$. (*c*) Indenter at slightly higher load suddenly penetrates. Contact pressure is about 80 kg mm$^{-2}$. Shape of indentation corresponds to an effective strain of about 20% (see figure 15).

due to the strengthening effect of the film but to its lubricating action. At the edges of the contact region very high stress-concentrations occur. If these cannot be relieved they will generate dislocations and plastic flow will take place without difficulty. The material will behave as though it contained dislocations and the hardness is of the same order as, though somewhat greater than, the 'ordinary' bulk hardness. This is the situation when the surfaces are clean. In the presence of a lubricating film these stresses are relieved and the metal is able to withstand very much higher stresses before flow occurs.

In another series of experiments Gane constructed a miniature form of the apparatus so that it could operate in the field of a transmission electron microscope. In most of these experiments a fine metal tip was pressed against a hard platten. Similar results were obtained. If the tip was fine enough the electron beam could penetrate the metal and show the generation of dislocations in the tip as 'blunting' proceeded.

In conventional microhardness measurements where a fine pointed indenter is used it is probable that the initial contact stresses before yielding occurs (in theory they would be infinite if the tip were perfectly sharp) are able to generate dislocations even in perfect portions of the crystal. Consequently the hardness will be comparable with that of the bulk material. It may be somewhat higher but not more than a factor of two or so. With

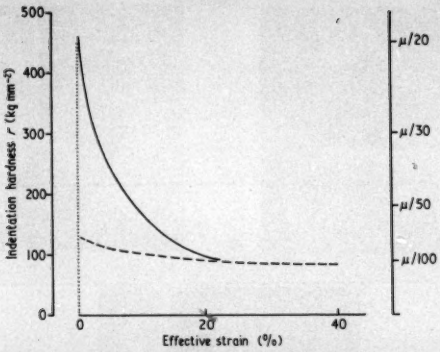



Figure 15. Indentation hardness of a single crystal of annealed gold as a function of strain introduced by the indentation process itself. Full line, when a surface film is present; broken line, in the absence of a surface film. The maximum hardness has a value of about 450 kg mm$^{-2}$ and resembles an upper-yield point: the bulk hardness in air is of order 30–40 kg mm$^{-2}$.

a rounded tip and a surface film which can relieve stress concentrations in the contact zone hardness values approaching the theoretical value for a perfect crystal may be observed.

## 4. Indentation hardness of elasto-plastic solids

### 4.1. *The basic concepts*

All the previous sections have ignored elastic effects in the indentation process. We now consider their role. In 1957 Samuels and Mulhearn observed that in some cases the deformation below a Vickers indenter did not show the outward flow of metal predicted by plasticity theory but resembled the deformation that would be expected if there was radial flow of the metal away from the indenter. In 1964 Marsh followed this up by suggesting that this was due to elastic yielding of the hinterland, the deformation resembling the expansion of a spherical cavity into a plastic–elastic solid by an internal (hydrostatic) pressure. If the indentation has a chordal diameter $d$ he assumed that there is a spherical core of diameter $d$ (radius $a=\frac{1}{2}d$) over which the indentation pressure $p$ acts. The plastic strains from this zone gradually diminish as one moves into the bulk of the specimen until they match the elastic strains in the hinterland at some radius $c$. This problem has already been analysed by Hill (1950) who showed that the pressure $p$ at which the cavity expands depends on the ratio of $Y$ to the elastic modulus. If $E$ is Young's modulus and $\nu$ Poisson's ratio he gives

$$\frac{p}{Y}=\frac{2}{3}\left(1+\ln\frac{E}{3(1-\nu)\,Y}\right). \tag{20}$$

Assuming for simplicity that $\nu=\frac{1}{2}$ this leads to

$$\frac{p}{Y}=0{\cdot}40+\frac{2}{3}\ln\frac{E}{Y}. \tag{21}$$

In a series of careful experiments with a variety of materials Marsh found that his hardness results followed a pattern very close to that of equation (21) but the constants were a little different. This may be because the deformation in hardness experiments is not exactly identical with the expansion of a hemispherical core but the better theory in §4.2 provides

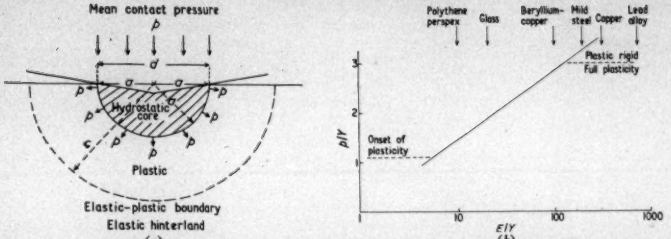



Figure 16. (*a*) The indentation process in an ideal elastic–plastic solid regarded as the expansion of a hemispherical core, (*b*) the variation of indentation pressure $p$ with the ratio $E/Y$ where $E$ is Young's modulus and $Y$ the yield stress. Results deduced by Marsh (1964) for a Vickers indenter. The horizontal broken lines also represent approximate limits for a spherical indenter.

a different explanation. For a Vickers indenter Marsh's results fitted the relation

$$\frac{p}{Y} \simeq 0{\cdot}07 + 0{\cdot}6 \ln \frac{E}{Y}. \tag{22}$$

The variation of $p/Y$ with the ratio $E/Y$ given by equation (22) is shown in figure 16(*b*). It is seen that when $E/Y$ is small, say 10, $p$ is only $1{\cdot}5Y$. This corresponds to a very 'elastic' material which is able to accommodate the plastic strains and so reduces the constraint on flow. This represents the behaviour of many plastics. For $E/Y \simeq 20$, $p = 1{\cdot}7$ to $2Y$: this corresponds to the behaviour of glasses. For $E/Y$ greater than about 150 the ratio $p/Y$ is greater than 3.

### 4.2. *The effect of indenter shape*

This analysis shows that for conical and pyramidal indenters of large apical angle the plastic yield pressure depends on the ratio $E/Y$. The behaviour resembles the radial expansion of a hemispherical core. If, however, the semi-angle of the cone or pyramid is less than about 50° the mode of deformation is different and is much closer to that of a plastic-rigid metal. The metal is 'cut' by the indenter and the displaced material is pushed up the sides of the indenter. For a spherical indenter the behaviour, as mentioned earlier, is different. The onset of permanent deformation occurs when $p \simeq 1{\cdot}1Y$. As the load is increased the size of the plastic indentation increases, the plastic zone grows until finally the whole of the material around the indentation is plastic.

A similar treatment to that given by Marsh has been described by Hirst and Howse (1969) for a wedge-shaped indenter. They treat the flow process as the radial expansion of a cylindrical core and obtain results resembling those of figure 16. The best analysis of the whole problem is due to Johnson (1970). He points out that the radial displacement of particles lying on the elastic–plastic boundary during an increment of penetration must accommodate the volume of material displaced by the indenter during this incremental movement. In this way he shows that the indentation process may be described in terms of the semi-angle $\theta$ of the indenter (or by the ratio $a/R$ for a spherical indenter). The parameters now plotted are $p/Y$ against $E \cot \theta / Y$ or $E(a/R)/Y$ and the results for pyramidal, conical and spherical indenters all lie on a single curve. For a conical or pyramidal indenter, equation (20) becomes

$$\frac{p}{Y} = \frac{2}{3}\left(1 + \ln \frac{E \cot \theta}{3Y}\right). \tag{23}$$

This is plotted in figure 17 and is probably the most useful indentation curve that has, as yet, been derived. It is interesting to note that this may also be used to describe the

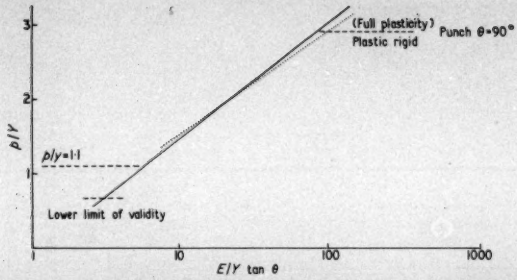



Figure 17. Variation of indentation pressure $p$ with the $E/Y \tan\theta$ for a conical indenter where $\theta$ is the semi-apical angle of the cone. Full line, theoretical curve due to Johnson (1970); dotted line, experimental results due to Marsh appropriately displaced (see figure 16(*b*)) on the assumption that a Vickers pyramid is roughly equivalent to a cone of $\theta = 70°$.

behaviour of a spherical indenter, if cot $\theta$ is replaced by $a/R$. For example it is seen that full plasticity ($p \simeq 3Y$) is reached for a value of the horizontal ordinate about 10 times greater than that at which the onset of plasticity occurs ($p \simeq 1{\cdot}1\,Y$). For a given material ($E/Y$ is a constant) and for an indenter of fixed radius $R$ this corresponds to a 10-fold increase in $a$, the radius of the indentation. This result agrees very well with the results already described in figure 5(*b*).

It is interesting to consider the limits of any hardness theory based on the idea of an expanding hemispherical core. If we go back to Hill's original analysis we find that the pressure in the core, when the plastic–elastic boundary is at distance $c$, is given by

$$\frac{p}{Y} = \frac{2}{3} + 2\ln\frac{c}{a}. \tag{24}$$

This implies that the elastic–plastic boundary coincides with the boundary of the plastic core ($c = a$) at $p = \frac{2}{3}Y$ and below this contact-pressure the analysis is no longer valid. In the region where $p \simeq 3Y$, $c \simeq 3{\cdot}2a$ and the elastic yielding of the hinterland apparently no longer influences the plastic flow of the material. The contact pressure now corresponds to the classical theory for a rigid-plastic solid.

## 5. Indentation hardness and creep of metals. Hot hardness

### 5.1. *Experimental procedure*

In industry it is often desirable to determine the strength properties of a new alloy or composite at elevated temperatures. The proper procedure is to prepare a tensile specimen and carry out a stress–strain experiment at a specified strain rate at a series of temperatures. Such procedures are time-consuming, involve complex apparatus and demand the fabrication of a relatively large specially shaped specimen. It is much easier to carry out a hardness measurement which demands simple apparatus and a small specimen with a flat polished surface. The hardness may be determined in a static experiment by loading the indenter and measuring the size of the indentation after the experiment is completed (Westbrook 1958). Alternatively (Fitzgerald 1963) a dynamic method may be used (see below): a hard sphere is dropped on to the specimen and its height of rebound determined.

Such measurements show that the hardness depends on the temperature: it also depends on the time of loading. Some results obtained under static loading conditions at a constant loading time are shown in figure 18 and it is seen that they fall into a simple pattern. At

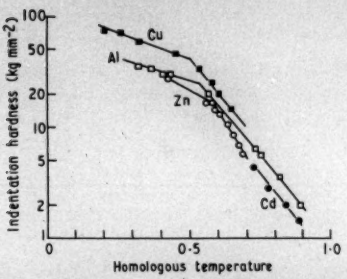



Figure 18. Indentation hardness $p$ as a function of homologous temperature $T/T_m$ for copper, aluminium and zinc. $T_m$, melting point; $T$, temperature of experiment on absolute scale. Constant time of loading. (From data by Westbrook 1959.)

temperatures below about $0{\cdot}57T_m$ (where $T_m$ is the absolute melting point) the hardness falls slightly with increasing temperature: above $0{\cdot}5T_m$ there is a much more rapid softening of the material. This is because self diffusion becomes much more important and the material shows relatively marked creep.

The static method is clearly more informative since the hardness diminishes with loading-time and this provides additional insight into the softening process. Atkins *et al.* (1966) have studied this in some detail. Indentation was carried out between crossed wedges or crossed cylinders of the material since this eliminated the problem of choosing an indenter hard enough and sufficiently stable chemically to withstand the very high temperatures used (up to 2000 °C). The apparatus was maintained in a vacuum of about $10^{-5}$ torr to reduce oxidation but an argon atmosphere would have been equally satisfactory.

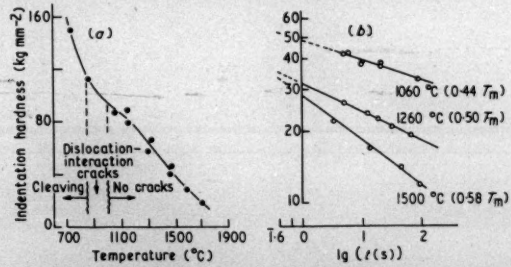



Figure 19. (*a*) Mutual indentation hardness of crossed wedges of single crystals of MgO as a function of temperature (constant loading time). Ductile behaviour is observed above about 1000 °C. (*b*) Mutual indentation hardness of single crystals of MgO in the ductile temperature range as a function of loading time $t$.

Some typical results are shown in figure 19 for single crystals of MgO. Below 900 °C the material behaved in a brittle manner; above 1000 °C it showed no cracking and indented in a ductile manner. Similar results were observed for polycrystalline $Si_3N_4$. In general single crystals of a given material behaved like polycrystalline specimens even if they contained some porosity; this is because, at higher temperatures, after compaction of the

porous material has occurred around the indentation self-diffusion dominates the flow properties (Atkins and Tabor, 1966).

### 5.2. *Hardness and creep above* $0{\cdot}5T_m$

In this temperature range, if tensile-creep experiments are carried out at constant stress $s$, the steady-state or viscous creep may be expressed by a relation of the form

$$\dot{\epsilon}_s = A_1 s^m \exp\left(\frac{-Q}{RT}\right) \tag{25}$$

where $m$ is of order 5 (Weertman 1955) and $Q$ is an appropriate activation energy.

The transient creep (Andrade 1910) which precedes this can be expressed by a relation derived by Mott of the form

$$\dot{\epsilon}_{tr} = A_2 \dot{\epsilon}_s^{1/3} t^{-2/3}. \tag{26}$$

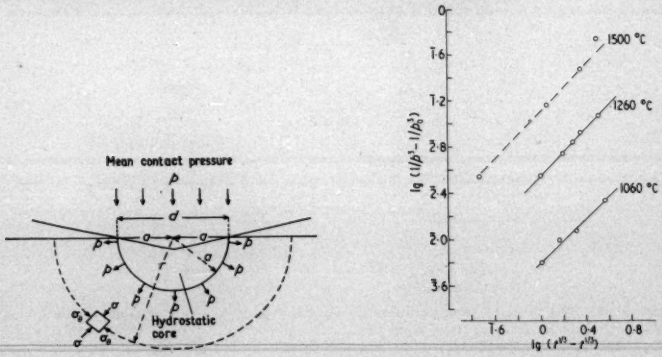
Combining this with equation (25) we obtain

$$\dot{\epsilon}_{tr} = A_3 s^{m/3} \exp\left(\frac{-Q}{3RT}\right) t^{-2/3} \tag{27}$$

or in terms of the shear strain $\gamma$ and the shear stress $\tau$

$$\dot{\gamma}_{tr} = A\tau^{m/3} \exp\left(-\frac{Q}{3RT}\right) t^{-2/3}. \tag{28}$$

All these creep equations have been derived from experiments where the stress is substantially constant, and not too large, and where every part of the specimen is undergoing homogeneous deformation. The situation is very different in a hardness indentation where stresses and strains are non-homogeneous and the driving pressure itself changes throughout the experiment. Nevertheless it may be shown that to a close approximation the transient creep equation may be applied to explain the indentation behaviour.




(*a*) (*b*)

Figure 20. (*a*) Indentation process under conditions of creep: a hemispherical core under the indenter is assumed to be under hydrostatic pressure equal to the indentation pressure. The growth of the indentation with time is assumed to be limited by the onset of transient creep at the outer boundary. (*b*) Plot of $p^{-1/3} - p_0^{-1/3}$ against $t^{-1/3} - t_0^{-1/3}$ on log–log ordinates for MgO (results taken from figure 19(*b*)). The theory predicts a slope of unity which is very close to the experimental results.

Following a suggestion by R. Hill, the indentation process is assumed to correspond to the plastic movement of a series of shells, concentric with the hemispherical core surrounding the indentation, into the bulk of the specimen (figure 20(*a*)). Unlike the Marsh treatment the elastic properties of the hinterland are ignored. Instead it is assumed that at some distance $r$ from the centre of the core the rate of flow of the material is determined by the transient creep of the material at this point. For an incompressible material it is easy to show that, if the radius of the indentation (this equals the radius of the core) is $a$, the shear strain rate at a distance $r$ is

$$\dot{\gamma}=\frac{3}{2}\frac{a^2}{r^3}\dot{a}. \tag{29}$$

If this is substituted into equation (28) it gives

$$\tau=K_1 a^{6/m}\dot{a}^{3/m}t^{2/m}r^{(-9/m-1)}\exp\left(\frac{-Q}{mRT}\right). \tag{30}$$

Now the plasticity equilibrium condition shows that at a distance $r$ the radial stress $\sigma$ is related to $\tau$ by the equation

$$\frac{\mathrm{d}\sigma}{\mathrm{d}r}=-\frac{4\tau}{r}. \tag{31}$$

From equation (30) we thus have an expression for $\mathrm{d}\sigma/\mathrm{d}r$ in terms of the size of the indentation $a$, $\dot{a}$ and the radius $r$. This may be integrated from $r=a$, at which region $\sigma$ equals the indentation pressure $p$, to $r=\infty$ at which region $\sigma=0$. We obtain

$$\sigma=p=K_2\left(\frac{\dot{a}}{a}\right)^{3/m}t^{2/m}\exp\left(\frac{Q}{mRT}\right). \tag{32}$$

Finally we note that $p$ is proportional to $W/a^2$ so that

$$a\propto W^{1/2}p^{-1/2}$$

$$\dot{a}\propto W^{1/2}p^{-3/2}\dot{p}.$$

Consequently

$$\frac{\dot{a}}{a}\propto\frac{\dot{p}}{p}. \tag{33}$$

Substituting in equation (32) and integrating for $p$ we obtain

$$p^{-m/3}-p_0^{-m/3}=A_4\exp\left(-\frac{Q}{3RT}\right)(t^{1/3}-t_0^{1/3}) \tag{34}$$

where $p$ is the hardness at time $t$ and $p_0$ the hardness at $t_0$ immediately after attaining the full load $W$.

For small values of stress ($s\leqslant 10^{-4}\times$ elastic modulus) the value of $m$ is near 5. For the stresses involved in hardness measurements ($s>10^{-3}\times$ elastic modulus) $m$ appears to be nearer 10. For simplicity we take a value $m=9$ and so write equation (34) as

$$p^{-1/3}-p_0^{-1/3}=A_4\exp\left(-\frac{Q}{3RT}\right)(t^{1/3}-t_0^{1/3}). \tag{35}$$†

If the results for MgO are replotted in this way we obtain a series of straight lines of constant slope equal to unity (figure 20(*b*)). The separation between the lines is a measure of $Q$; the value is 110 kcal mol$^{-1}$. Similar data obtained for other materials (Atkins *et al.* 1966) are shown in table 3. (For a more recent hardness–temperature study of $TiC_x$ see Samsonov *et al.* (1970): they use a rather unusual creep relation.)

† It is interesting to note that if creep is primarily viscous (as it may well be at $T>\frac{1}{2}T_m$) so that equation (25) is used in the analysis, the final result is identical with equation (35).

**Table 3. Activation energies $Q$ derived from hardness and self diffusion**

| Material | Homologous temperature range | $Q$ (kcal mol$^{-1}$) | $Q$ for self diffusion (kcal mol$^{-1}$) |
|---|---|---|---|
| MgO | 0·3–0·5 | 30 | — |
| | 0·5–0·7 | 110 | 105[a] |
| $Si_3N_4$ | 0·5 | 30–40 | — |
| TiC | 0·3–0·5 | 30 | — |
| | 0·5–0·7 | — | 150[b] |
| SiC | 0·5 | 45 | — |
| WC | 0·3–0·5 | 20–40 | — |
| | 0·5–0·7 | 110 | — |

[a] Groves and Kelly (1963) [b] Hollox and Smallman (1964)

Recently K. L. Johnson (unpublished) has provided a more rigorous analysis which enables him to include in the final relation a factor to describe the shape of the indenter. It turns out that for a pyramidal or conical indenter of semi-angle $\theta$ the contact pressure is only changed by a factor $(\cot\theta)^{1/9}$. This is consistent with experiments which show that the hardness and creep behaviour depend little on the indenter shape.

It is evident from the analysis and from the experimental results that indentation hardness measurements may be used to study the yield and creep properties of solids particularly at elevated temperatures. One set of experiments of particular interest is that showing the variation in hardness with temperature of WC and TiC. There is a catastrophic fall in hardness at temperatures well below $0{\cdot}5T_m$; at 1200 °C (about $0{\cdot}5T_m$) the hardness has fallen from its room temperature value of over 2500 kg mm$^{-2}$ to less than 100 kg mm$^{-2}$ (figure 21). Since these materials are often used as cutting tools in the (unlubricated)

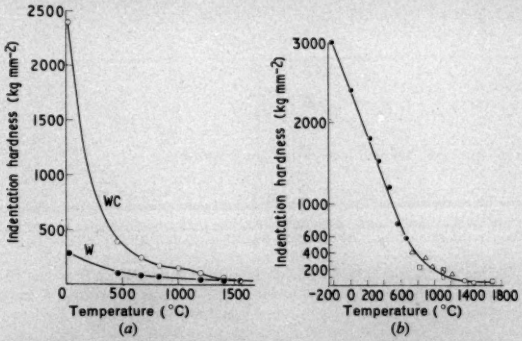



Figure 21. Variation of hardness with temperature of WC and TiC.

machining of hard metals where very high surface temperatures are developed it would be interesting to know if indeed such softening occurs. It may be remarked that Hollox (1968) has found that the high-temperature softening of TiC is greatly diminished if it is converted into a mixed carbide containing VC.

## 6. Elastic recovery of hardness indentations: rebound hardness

### 6.1. *Recovery of indentations formed by spherical indenters: 'shallowing'*

When a spherical indenter is pressed into a metal surface and then removed, there is some elastic recovery of the indentation. Careful measurements show that there is little change

in the chordal diameter $d_0$ of the indentation but a marked increase in its curvature. The indentation is shallower than if it were of the same curvature as the indenter. To a close approximation the recovered indentation is spherical in shape. Consequently we may calculate its radius of curvature $r_2$ in terms of the radius of the indenter $r_1$ and the elastic properties of the solids (figure 22). For if we were to reapply the original load $W$ the surfaces would deform elastically, according to Hertz's equation for the deformation of contacting spherical surfaces, until contact just occurred over the diameter $d_0=2a_0$. According to Hertz the relation is

$$\frac{1}{r_1}-\frac{1}{r_2}=\frac{3Wg}{4a_0^3}\frac{1}{E'} \tag{36}$$

where

$$\frac{1}{E'}=\frac{1-\nu_1^2}{E_1}+\frac{1+\nu_2^2}{E_2} \tag{37}$$

where $E_1$, $E_2$ are Young's modulus and $\nu_1$, $\nu_2$ Poisson's ratio of the indenter and metal respectively.

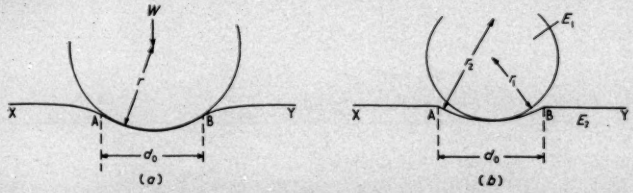



Figure 22. (*a*) Indentation produced by a spherical indenter in a plastic–elastic solid, (*b*) shape of recovered indentation after indenter is removed.

Of course there are residual elastic stresses left in the material surrounding the indentation. Further, the recovered shape is not *exactly* spherical so that the Hertzian equation cannot be applied rigorously. Nevertheless experiments with a wide range of metals show that shallowing of spherical indentations follows equation (36) rather well.

### 6.2. *Rebound hardness for spherical indenters*

If a spherical indenter of mass $m$, radius $r_1$ is dropped from a height $h_1$ on to a massive metal specimen it produces a plastic indentation of diameter $d_0$ ($=2a_0$) and rebounds to a height $h_2$. If it be assumed that a mean dynamic indentation pressure $p$ exists during indentation the plastic work involved is $pv$ where $v$ is the volume of the plastic dent, that is, of the recovered indentation. It is reasonable to assume that the energy of rebound arises from elastic recovery of the indentation (and the indenter) after plastic deformation has come to an end. This may be estimated by considering the reverse process, that is the elastic work done when the indenter is replaced in the recovered indentation and a force applied for a short time sufficient to deform the surfaces until the indenter just touches over the circle of contact of radius $a_0$. Supposing at any stage of this process the force is $F$ when the circle of contact is $a$ ($a<a_0$). According to equation (36) we may write

$$F=\frac{4a^3}{3}\left(\frac{1}{r_1}-\frac{1}{r_2}\right)\left(\frac{1}{E'}\right)^{-1}. \tag{38}$$

At this stage the surfaces will have sunk together by a distance $x$ given according to Hertz by

$$x=a^2\left(\frac{1}{r_1}-\frac{1}{r_2}\right). \tag{39}$$

So that

$$dx = 2a\left(\frac{1}{r_1} - \frac{1}{r_2}\right) da. \tag{40}$$

The work done by $F$ in moving $dx$ is $F\,dx$. By integration and rearranging the result we find that the work done in increasing $F$ from zero to its final value $F_0$ when $a = a_0$ is $\frac{3}{10}(F_0^2/a_0)(1/E')$ or $\frac{3}{8}(1/r_1 - 1/r_2)F_0 a_0^2$.

This may be equated to the energy of rebound so that

$$mgh_2 = \frac{3}{10}\frac{F_0^2}{a_0}\frac{1}{E'}. \tag{41}$$

The volume $v$ of the recovered indentation is approximately $\pi a_0^4/4r_2$ whilst the volume the indentation would have if it possessed the same curvature as the indenter is $V = \pi a_0^4/4r_1$. Combining equations (38) and (41) and rearranging and noting that at the end of the process $F_0 = p\pi a_0^2$ we get

$$mgh_2 = \frac{2}{5}\left(\frac{1}{r_1} - \frac{1}{r_2}\right) p\pi a_0^4 = \tfrac{5}{8}(pV - pv) \tag{42}$$

or

$$\tfrac{5}{8}mgh_2 = pV - pv. \tag{43}$$

The energy expended in forming the indentation is

$$pv = mgh_1 - mgh_2. \tag{44}$$

Combining this with equation (43) we obtain

$$p = mg\frac{h_1 - \frac{5}{8}h_2}{V}. \tag{45}$$

The recovered indentation radius is thus eliminated and we have a simple expression for $p$ in terms of the indenter radius and the height of impact and rebound. In general $p$ calculated in this way is a little bigger than the static value. But for soft metals the difference can be very marked. The reason, of course, is the short time of loading. To a good approximation the duration of the plastic indentation is independent of the velocity of impact and is given by

$$t = \frac{\pi}{2}\left(\frac{m}{2\pi pr}\right)^{1/2} \tag{46}$$

so that for typical experiments $t$ is of the order $10^{-4}$ to $10^{-5}$ s. For lead at room temperature this short-time loading increases the yield pressure by about 50%. But at a temperature of 100 °C or more (this is above $0{\cdot}5T_m$) the creep rate is much more rapid. As equation (32) indicates if the loading time is changed from say $10^0$–$10^{-5}$ s (i.e. a factor of $10^{-7}$) the yield pressure will change by a factor of 5 to 10. Experimentally one finds a factor of 9:1 for the ratio of the dynamic to the static hardness of lead at 100 °C.

Finally the equation (45) may be combined with equation (41) to eliminate the volume $V$ of the indentation: $p$ is then expressed solely in terms of the elastic properties.

$$p^5 = \frac{h_2^4}{(h_1 - \frac{5}{8}h_2)^3}\frac{mg10^4}{3^4 \times 4^3\pi^5 r_1^3}\frac{1}{(1/E')^4}. \tag{47}$$

We note that if $h_1 = h_2 = h$ we obtain

$$p^5 = \frac{640}{3^4\pi^5}\frac{mgh}{r_1^3}\frac{1}{(1/E')^4}. \tag{48}$$

This is precisely the value given by Hertz for the mean pressure achieved in a purely elastic collision. Equation (47) thus covers the whole range of collision conditions. It explains the calibration characteristics of the Shore Rebound Scleroscope.

6.3. *Rebound for conical indenters*

A similar study has been described for conical indenters (semi-apical angle $\theta$) by Stilwell and Tabor (1961). It shows that to a first approximation a conical indentation recovers elastically to give a cone of wider angle than that of the indenter (figure 23(*a*)). Analysis

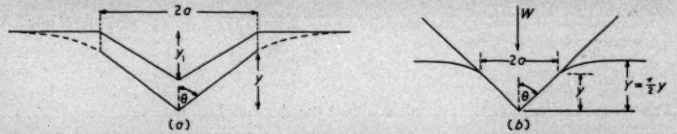



Figure 23. Indentation produced (*a*) in an elastic–plastic solid by a rigid conical indenter. When the indenter is withdrawn the indentation recovers elastically to give a shallower conical impression ($y_1 < y$). (*b*) In an elastic solid the total depth of penetration $Y=(\pi/2)\,y$.

shows that this may be described in terms of standard elastic equations and the pressure during indentation may be calculated to give

$$p=\frac{mg\{h_1-(1-2/\pi)\,h_2\}}{V} \tag{49}$$

or, eliminating the volume of the indentation

$$p=\frac{h_2}{\{h_1-(1-2/\pi)\,h_2\}}\,\frac{\cot\theta}{\pi(1/E')}. \tag{50}$$

One should note once again that the recovery is complicated by residual stresses so that the analysis carries some limitations. A study of the shallowing of wedge-shaped indentations has recently been published by Hirst and Howse (1969).

**7. Indentation hardness of polymers and rubbers**

7.1. *Polymers*

Thermoplastics such as nylon, perspex, polythene have very complex deformation properties. For small strains they deform elastically, for larger strains they flow irreversibly. Both the elastic modulus and the yield stress depend on the rate of deformation so that they are generally described as visco-elastic solids. Nevertheless to a first approximation they behave like metals and can be characterized by an elastic modulus $E$ and a yield stress $Y$. In addition under conditions of complex stress they appear to follow a Tresca or von Mises yield criterion. We should therefore expect their indentation behaviour to resemble that of metals. However, for most thermoplastic polymers the ratio $E/Y$ is of order 10 compared with a value of 100–1000 for metals. Consequently, as described in a previous section, elastic yielding of the hinterland is more important, there is less constraint on plastic flow and the hardness is appreciably less than $3Y$. The behaviour will also depend on the shape of the indenter. For the Vickers indenter the indentation pressure turns out to be of order $1{\cdot}5\,Y$ so that

$$\mathrm{DPH}\simeq 1{\cdot}4\,Y.$$

Vickers indentations in polymers usually show a marked pincushion appearance. Consequently an accurate measurement of the true indentation area from the diagonals is not generally possible. Further, there is elastic recovery when the indenter is removed and, with time, as a result of relaxation the indentation may actually disappear. Hirst and

Howse (1969) have discussed in detail the indentation and elastic recovery of polymers indented by two-dimensional wedges. They have not, however, considered their visco-elastic properties. This has been done in a very elegant way both for wedges and for axially symmetric indenters by K. L. Johnson (unpublished).

### 7.2. *Rubbers*

The hardness of rubbers is usually measured by a penetrometer. A conical indenter is pressed into the surface of the rubber specimen and the depth of penetration recorded on an appropriate scale. In this it resembles the Rockwell hardness test. The penetration is determined mainly by the elastic modulus of the rubber. For a cone of semi-apical angle $\theta$ the average indentation pressure is

$$p = \frac{E \cot \theta}{2(1-\nu^2)} \tag{51}$$

where $\nu$ is Poisson's ratio. For most rubbers this is about $\frac{1}{2}$, so that

$$p \approx \tfrac{2}{3} E \cot \theta. \tag{52}$$

It may be noted (figure 23(*b*)) that if the conical indentation has height $y$, the total depth of penetration of the indenter is $\frac{1}{2}\pi y$.

## 8. Scratch hardness of solids

### 8.1. *The Mohs scale*

It has long been the tradition amongst mineralogists and lapidaries to assess the hardness of stones or minerals by some type of scratch test. It was first put on a quantitative basis by the Austrian mineralogist Mohs (1824). He proposed a scale of ten minerals such that each mineral will scratch the one on the scale below it, but will not scratch the one above it. Mohs was fully aware of the difficulty of constructing such a scale and in selecting his standard minerals he emphasizes that: 'The intervals between every two members of the scale be not so disproportionate as either to render its employment more difficult or to hinder it altogether.' With these precautions in mind Mohs finally proposed as his ten basic minerals:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1. | talc | 2. | gypsum | 3. | calcite | 4. | fluorite | 5. | apatite |
| 6. | orthoclase | 7. | quartz | 8. | topaz | 9. | corundum | 10. | diamond |

He realized that the gap between corundum and diamond was larger than it should have been but seemed satisfied with the remainder of the scale. However, Mohs nowhere explains his criterion of equality of intervals and at first sight it would seem that such a scale might be so arbitrary as to have no basic physical significance. As we shall see, there is a sound physical basis on which equal intervals can be constructed and the Mohs scale follows it surprisingly well (Tabor 1954).

### 8.2. *Indentation hardness of minerals*

In discussing the scratch-hardness of minerals it might seem that we are concerned primarily with the behaviour and physical properties of relatively brittle materials. This, however, is not so. The work of Bridgman and others has shown that under sufficiently high hydrostatic pressures brittle materials may be prevented from fracturing, so that any deformation which they undergo under these conditions is essentially plastic. As we saw earlier, the stresses around an indenter are equivalent to a hydrostatic pressure on which is superposed a shear stress. With many materials, these hydrostatic pressures are sufficient to inhibit brittle fracture, and very satisfactory plastic indentations may be obtained even though some cracking may occur. Further, in the scratch process itself, the conditions at the contact region are similar to those around a static indenter. Here again a detailed examination shows that, although some fragmentation may occur, the deformation is

dominated by the plastic flow of the material. Since, therefore, both the scratching process and the static indentation process are determined primarily by the plastic properties of the material, we may expect to find some correlation between the Mohs hardness and the indentation hardness. This correlation does in fact occur, as is shown in figure 24, the data being based on indentation hardness measurements made with the standard Vickers indenter and also with the Knoop type of indenter which appears to give somewhat less cracking of brittle materials. The general trend is clear. The indentation hardness rises monotonically with increasing steps for each increment of the Mohs scale.

### 8.3. *The scratch-hardness of metals*

Since scratch-hardness involves, primarily, the plastic properties of minerals, its investigation is greatly simplified by using metals instead of minerals. A simple experiment along the following lines at once reveals the basic characteristic of a scratch-hardness scale. By suitable heat treatment, a strip of metal is rendered soft at one end and hard at the other, with a fairly uniform increase in indentation hardness along its length. Another metal specimen of uniform intermediate hardness is prepared with a sharp point at one end, and the point is dragged over the strip from the soft to the hard end. If the surfaces are lubricated it is then found that, over the soft portion of the strip, the friction is high and the motion intermittent, and a fine chip or shaving is produced. The behaviour remains much the same as the point approaches the harder end, until at a critical hardness of the surface the friction suddenly drops to a low value and the surface damage becomes negligible. Scratching has ceased. Applying this experimental procedure as a general criterion of scratching, it is found that a point of indentation hardness $H_p$ will scratch a surface of indentation hardness $H_s$ only if $H_p \geqslant 1{\cdot}2H_s$. The reason for this is not clear, for it does not appear to depend very critically on the shape of the point. Assuming, however, that scratching just occurs when $H_p = 1{\cdot}2H_s$, we may construct a hardness scale in which every unit is 1·2 times as hard as the preceding one. The scratch-hardness number $M$ is then related to the indentation hardness $H$ by a relation of the form

$$H = k(1{\cdot}2)^M.$$

Taking logarithms of both sides we have

$$\log H = M \ln 1{\cdot}2 + \text{constant}. \tag{53}$$

Thus a plot of ln $H$ against $M$ should give a straight line of slope 1·2. In practice we may wish to avoid the chance of an overlapping of the units by increasing the ratio to something greater than 1·2. Provided this new ratio is kept constant the general characteristic will still be a linear relation between ln $H$ and $M$.

### 8.4. *Mohs hardness and indentation hardness of minerals*

The results obtained by Winchell and by Krushchov and by Taylor, shown in figure 24, have been replotted in figure 25 as ln $H$ against $M$. It is seen that, if diamond be excluded, the relation ln $H = kM$ is surprisingly well obeyed, the value of $k$ being roughly ln 1·6 instead of ln 1·2.† This suggests that each interval on the Mohs scratch-hardness scale could be subdivided into two or three finer steps: but a finer subdivision would not be possible since it would involve steps in which the hardness difference ($H_p < 1{\cdot}2H_s$) would not be sufficient to produce well-defined scratching.

The value of $k$ on the Mohs hardness scale corresponds to a ratio of indentation hardness for each Mohs unit of about 1·6. Thus the Mohs hardness scale gives scratch-hardness values which correspond to fairly well-defined indentation-hardness values, each increment on the Mohs scale corresponding roughly to a 60% increase in indentation hardness. It is

† An alternative relation $M = kH^{1/3}$ has been suggested by Krushchov (1949). Although this fits the data satisfactorily its physical significance is obscure.

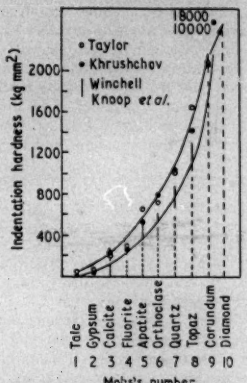



Figure 24. Relation between indentation hardness values and Mohs hardness numbers. Vickers indenter ○ ●; Knoop indenter |.

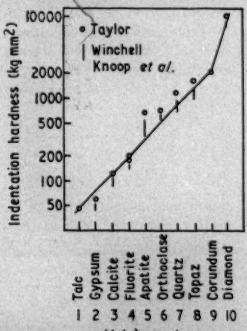



Figure 25. Relation between Mohs' hardness number $M$ and the logarithm of the indentation hardness $H$. Excluding diamond it is seen that $M$ is roughly proportional to ln $H$. Each Mohs interval corresponds to an increase in hardness value by a factor of about 1·6 (Tabor 1954).

clear that Mohs did not simply choose ten common minerals arranged in order of increasing hardness. It would seem that he experimented with a much larger number until he had satisfied himself that he had indeed obtained 'equality of the intervals'.

### 8.5. *Effect of adsorbed films on the hardness of minerals*

Adsorbed films of water and other surface active materials have little effect on the hardness of metals. They may have a very large influence on the hardness of minerals. Westbrook and his colleagues have studied this in some detail, in particular the effect of water vapour on materials such as GaSb, CdS, ZnO, SiC, diamond, LiF, quartz and glass.

The surfaces are first heated in argon at 250 °C to 300 °C, cooled and then immersed in toluene. Microhardness measurements carried out under toluene correspond to 'dry' conditions. If the surfaces are then exposed to water vapour or immersed in water the indentation hardness falls by between 10 and 20% depending on the material. With crystalline solids such as LiF the surface may be etched to show a rosette of dislocations around the indentation. These are found to be appreciably larger relative to the size of the indentation on wet surfaces. It is suggested that the effect is due to interaction between the adsorbate and point defects. On this view (Westwood *et al.* 1967) dislocation motion in non-metals is affected by electronic and strain interactions more significantly than by the Peierls resistance. Non-polar materials have little effect on the hardness but dimethyl sulphoxide (DMSO) actually increases the hardness in some cases. Crystals with non-centrosymmetric structures such as GaSb, CdS exhibit different behaviour on the opposite polar faces of the crystal. When dry both faces have the same hardness; when wet the B face shows a larger drop in hardness than the A face probably because adsorption of water occurs more readily on this face.

With amorphous materials the dislocation–adsorbate interaction is hardly applicable. Westbrook suggests that there may be some bond-strength weakening at the surface, or that the liquid film reduces the friction between the indenter and solid so reducing the hardness. The most plausible explanation (in the reviewer's opinion) is the formation of a softer silicacious layer.

The drop in hardness in all cases is limited to a very thin surface layer. Most of the measurements were carried out at loads which gave indentation depths below 3 micrometres. If the load was increased to give a greater depth of penetration the hardness increased. For example with InSb the low 'wet' hardness (about 100 kg mm$^{-2}$) was observed for a depth of 1 micrometre. For a depth of 5 micrometres the hardness was the same as for a 'dry' surface (about 170 kg mm$^{-2}$).

These results emphasize the care that must be taken to control the nature of adsorbed films when microhardness measurements are made on non-metallic solids.

## 9. Conclusion

It is evident from this review that the indentation hardness or yield pressure $p$ of metals is essentially a measure of their plastic properties. For a wide range of metals $p \simeq 3Y$ where $Y$ is the uniaxial flow stress of the metal. With pyramidal or conical indenters for indentations above some small critical size the hardness, because of geometric similarity, is independent of the size of the indentation and hence independent of the load. With spherical indenters this is not so, the hardness increasing with size of indentation. From the increase in hardness with load a semi-quantitative estimate may be made of the work-hardening characteristics of the metal. Similarly, from the variation of hardness with temperature and time of loading, information concerning the creep properties of the metal may be obtained.

If the material shows appreciable elasticity ($E/Y$ small) the yielding of the elastic hinterland may impose less constraint on plastic flow and the factor of proportionality between $p$ and $Y$ will be less than 3. This is just detectable with very hard metals. It is much more marked for thermo-plastic polymers ($p \simeq 1{\cdot}5Y$) and for glasses ($p \simeq 2Y$). In the limiting case where rubber-like materials are involved the deformation is elastic and the indentation is primarily a measure of the elastic modulus.

In the indentation and scratching of brittle solids such as minerals the high hydrostatic pressures developed around the deformed region are often sufficient to inhibit brittle fracture. Under these conditions the deformation is primarily plastic. For this reason there is fairly good correlation between indentation and scratch hardness since both are essentially a measure of the plastic and not the brittle properties of the solid. This provides a simple physical basis for the Mohs scratch-hardness scale.

Finally we may make a few comments on the relation between hardness and molecular structure. For metals the structural factors affecting hardness are the same as those

affecting the ease or difficulty of producing plastic flow. Strongly bonded metals (those with high melting points) will generally be harder than low melting metals. For example the hardness of Titanium is 200 times that of Indium. Dislocation entanglements such as those which occur when the metal is deformed restrict the ease of dislocation movement and the hardness may increase several fold (work-hardening). The greatest effect is produced by alloying which pins dislocations and can give very high hardness values (beryllium–copper 360 kg $mm^{-2}$ compared with 40 for copper; ball race steel 900 kg $mm^{-2}$ compared with 70 for iron). With single crystals the indentation hardness is influenced by the distribution of significant shear stresses on a limited number of slip planes and this accounts for anisotropic hardness behaviour. On the other hand, if indentations are carried out at extremely small loads, the volume deformed may contain no dislocations and the hardness may approach that of the perfect crystal.

With polymers and rubber-like materials the elastic and plastic properties depend on the chain length, the degree of cross-linking and the crystallinity of the material. Short chains are associated with low modulus and low hardness: long chains with increased modulus and hardness. The greatest hardness is observed with heavily cross linked molecules, as in ebonite.

The hardness of minerals depends very markedly on the type of bonding. Minerals in which the molecular chains or molecular sheets are attracted to their neighbours by van der Waals force are relatively weak and soft. Typical examples are talc and gypsum which have a Mohs hardness of 1 and 2 respectively. With ionic crystals the bonding is stronger but the bonding forces are electrostatic and are not directional. Such solids tend to acquire a crystal structure determined largely by geometric considerations of packing consistent with electric neutrality. Deformation is not particularly difficult; typical Mohs' hardness values being about 2 for NaCl, 4 for $CaF_2$ (fluorite), and 5 for $Ca_{10}(Po_4)_6F_2$ (fluorapatite). The equivalent indentation hardness values are approximately 30, 180, 400 kg $mm^{-2}$. As the degree of covalent bonding increases the bonding strength increases and in addition becomes strongly directional. For such materials it may become difficult to start the movement of dislocations (Peierls–Nabarro force) and the resistance to plastic flow is very high. The most marked examples are diamond and boron nitride which are the hardest known solids possessing a Mohs hardness 10, this is equivalent to an indentation hardness greater than 10 000 kg $mm^{-2}$.

**Acknowledgments**

I wish to express my thanks to Dr N. Gane, Dr M. J. Murray, Dr C. A. Brookes, Dr D. L. Kohlstedt, Mr R. H. J. Hannink and Mr P. Pfaelzer for helpful criticisms of the text I am particularly grateful to Dr K. L. Johnson for showing me unpublished chapters of a book on contact stresses which he is currently preparing for publication.

**Select bibliography**

*Books and survey articles*

BÜCKLE, H., 1959, *Progress in Micro-Indentation Hardness Testing, Metal. Rev.* (London: Inst. Metals), **4**, 49.

LYSAGHT, V. E., 1949, *Indentation Hardness Testing* (New York: Rheinhold).

MOTT, B. W., 1957, *Micro-indentation Hardness Testing* (London: Butterworths).

O'NEILL, H., 1934, *The Hardness of Metals and Its Measurement* (London: Chapman and Hall).

—— 1967, *Hardness Measurement of Metals and Alloys* (London: Chapman and Hall).

SMALL, L., 1966, *Hardness: Theory and Practice* (Michigan: Ferndale).

TABOR, D., 1951, *The Hardness of Metals* (Oxford: Clarendon Press).

WEINGRABER, H. VON, 1952, *Technische Härtemessung* (Munich: Carl Hanser Verlag).

*Papers*

ANDRADE, E. N. DA C., 1910, *Proc. R. Soc.* A, **84**, 1.

ATKINS, A. G., SILVERIO, A. A., and TABOR, D., 1966, *J. Inst. Metals*, **94**, 369.

ATKINS, A. G., and TABOR, D., 1966, *Proc. R. Soc.* A, **292**, 441.

BABYAK, W. J., and RHINES, F. N., 1960, *Trans. Metal. Soc. AIME*, **218**, 542.
BASSETT, W. H., and DAVIS, C. H., 1919, *Trans. Metal. Soc. AIME*, **60**, 428.
BERGSMANN, E. B., 1944, *Tek. Tidskr.*, **74**, 1297.
BIERBAUM, C., 1920, *Iron Age*, **106**, 1.
BOWDEN, F. P., and TABOR, D., 1964, *Friction and Lubrication of Solids* (Oxford: Clarendon Press), Part II.
BRINELL, J. A., 1920, *2nd Cong. Int. Méthodes d'Essai, Paris.* (For first English account see Wahlberg, A., 1901, *J. Iron and Steel Inst.* **59**, 243.)
BROOKES, C. A., O'NEILL, J. B., and REDFERN, B. A. W., *Anisotropy in the Hardness of Single Crystals. Proc. R. Soc.*, in the press.
COURTEL, R., 1961, *Comp. Rend.*, **253**, 1758.
DANIELS, F. W., and DUNN, C. G., 1949, *Trans. Am. Ceram. Soc.*, **41**, 419.
DUGDALE, D. S., 1958, *J. mech. Phys. Solids*, **6**, 85.
DYER, L. D., 1961, *Acta Met.*, **9**, 928.
FITZGERALD, L. M., 1963, *J. less-common Metals*, **5**, 356.
GANE, N., 1970, *Proc. R. Soc.* A, **317**, 367.
GARFINKLE, M., and GARLICK, R. G., 1968, *Trans. Metal. Soc. AIME*, **242**, 809.
GROVES, G. W., and KELLY, A., 1963, *J. appl. Phys.*, **34**, 3104.
GRUNSWEIG, J., LONGMAN, I. M., and PETCH, N. J., 1954, *J. mech. Phys. Solids*, **2**, 81.
HALL, E. O., 1951, *Proc. Phys. Soc.* B, **64**, 747.
—— 1954, *Nature*, **173**, 948.
HANNEMAN, R. E., and WESTBROOK, J. H., 1968, *Phil. Mag.*, **18**, 73.
HILL, R., 1950, *The Mathematical Theory of Plasticity* (Oxford: Clarendon Press).
HIRST, W., and HOWSE, M. G. J. W., 1969, *Proc. R. Soc.* A, **311**, 429.
HOLLOX, G. E., 1968, *Mat. Sci. Engng*, **3**, 121.
HOLLOX, F. E., and SMALLMAN, R. E., 1964, *Proc. Br. Ceram. Soc.*, **1**, 211.
JINDALL, P. C., and ARMSTRONG, R. W., 1967, *Trans. Metal. Soc. AIME*, **239**, 1856.
—— 1969, *Trans. Metal. Soc. AIME*, **245**, 623.
JOHNSON, K. L., 1970, *J. mech. Phys. Solids*, **18**, 115.
KHRUSCHOV, M. M., 1949, *Zavod. Labor.*, **15**, 243.
KLEESATTEL, C., 1966, *J. Acoust. Soc. Am.*, **39**, 404.
KNOOP, F., PETERS, C. G., and EMERSON, W. B., 1939, *J. Res. Nat. Bur. Stand.*, **23**, 39.
MARSH, D. M., 1964, *Proc. R. Soc.* A, **279**, 420.
MEYER, E. Z., 1908, *Ver. Deut. Ing.*, **52**, 645.
MOHS, F., 1824, *Grundriss der Mineralogie* (English translation by Haidinger, W., 1825, *Treatise on Mineralogy* (Edinburgh: Constable)).
PARTRIDGE, P. G., and ROBERTS, E., 1963, *J. Inst. Met.*, **92**, 50.
PETCH, N. J., 1953, *J. Iron Steel Inst.*, **174**, 25.
SAMSONOV, G. V., KOVALCHENKO M. S., DZEMELINSKII and UPADYAYA, G. S., 1970, *Phys. Stat. Solidi*, **1**, 327
SMITH, R., and SANDLAND, G., 1925, *J. Iron. Steel Inst.*, **1**, 285.
STILWELL, N. A., and TABOR, D., 1961, *Proc. Phys. Soc.*, **78**, 169.
TABOR, D., 1954, *Proc. Phys. Soc.* B, **67**, 249.
—— 1948, *Proc. R. Soc.* A, **192**, 247.
TAPLIN, D. M. R., 1966, *J. Nucl. Mater.*, **19**, 208.
TAYLOR, E. W., 1949, *Miner. Mag.*, **28**, 718.
VOCE, E., 1951, *J. Inst. Metals*, **79**, 465.
WEERTMAN, J., 1955, *J. appl. Phys.*, **26**, 1213.
WESTBROOK, J. H., 1957, *Proc. Am. Soc. Test. Mater.*, **57**, 873.
WESTBROOK, J. H., and JORGENSEN, P. J., 1968, *Am. Mineralogist*, **53**, 1899.
WESTWOOD, A. R. C., GOLDHEIM, D. L., and PUGH, E. N., 1967, *Phil. Mag.*, **15**, 105.

E. R. PIKE

Physics Group, Royal Radar Establishment, Malvern, Worcestershire

# Optical spectroscopy in the frequency range 1–$10^8$ Hz†

**Abstract.** A review is given of recent developments in optical spectroscopy in the resolution range $10^6$ to $10^{14}$. Both statistically-based intensity-fluctuation analysis and heterodyne methods are discussed.

## 1. Introduction

The conceptual possibilities accompanying the advent of the laser of utilizing a range of existing radio and microwave techniques in the infrared and optical regions of the spectrum were thwarted to a large extent in the early years by a number of technical factors. The first lasers were by no means the equivalent of the klystron or the radio-frequency oscillator in either stability or spectral purity, unless one remained content with extremely low powers. As the years have passed, however, one has seen a gradual infiltration, in many areas of optics, of well-known microwave methods. At the present time one can find in use systems employing heterodyne, homodyne and intensity-fluctuation detection, in addition to the high-$Q$ optical-frequency filter, and we can now regard all this activity as the opening up of a new field of optical spectroscopy in the range 1–$10^8$ Hz which was previously quite inaccessible to optical investigation. Increasingly optical methods have offered new approaches to problems both technological and fundamental, and in some cases these have superceded in speed and accuracy more conventional instrumentation.

In this paper recent progress in the subject will be reviewed. Although, as we shall discuss, the methods are applicable, in principle, to the spectral analysis of arbitrary light fields, the applications to date have been connected with laser light and laser scattering experiments.

Some of the most interesting of these are characterized by very weak optical signals where the quantal nature of the detection process is a dominant factor. The signals to be processed consist then of a train of pulses representing photodetections and the spectral information lies in the density of the pulses as a function of time. We shall be concerned therefore in this paper with the spectral analysis of the fluctuations of numbers of detected photons over specified time intervals. Direct measurement of temporal correlation has been used recently in preference to harmonic analysis and clipping methods of microwave technique have been transposed into the optical domain. We also discuss some experiments at 10 $\mu$m where heterodyne techniques help overcome noise problems. Reviews are given by Benedek (1968), Pike (1969) and Cummins and Swinney (1970).

A number of diverse applications will be illustrated including the remote detection of wind velocity, the study of turbulence in liquids, the measurement of diffusion constants of protein and polymer molecules, the measurement of thermal transport in liquids and the study of sound waves in solids.

In the next section we discuss methods of detection of the optical signal, taking regard of the difficulty that the optical frequency is too high to follow. In §3 the question of post-detection processing is taken up, with a comparison of the methods available. In §4 the results of recent calculations of performance of correlator systems are presented and the last section is devoted to applications.

## 2. Detection methods

### 2.1. *Straight detection*

To be completely general we should begin by considering an optical field with arbitrary spectrum falling upon a detector. In the spirit of my introduction we might imagine this

† This paper was read, by invitation, at the International Quantum Electronics Conference, Kyoto, Japan, September 1970.

incident field to be the analogue in the optical domain of the signals, say, which are picked up by a radio-frequency antenna. Now here we come to the first difference between the optical and microwave or radio-frequency regions. It is, in principle, possible, although nowadays not often done, to filter a radio-frequency signal into its separate spectral components immediately following the detector, which in the radio-frequency case is the antenna, and to process them at this frequency until demodulation. Because of the much higher frequencies of the optical signals and the inability of even the very fastest of electronic devices to follow the optical frequency, such a straight detection is not possible in the optical region unless one places one's filter before the optical detector. Such an instrument would, for instance, be typified by a scanning Fabry–Perot interferometer where the étalon would correspond to a tuning filter. The following detector then acts as a rectifying demodulator and we have the analogue of a straight radio receiver. This is classical optical spectroscopy and we discuss below newer methods with vastly superior resolution which make use of the coherence properties of laser light (heterodyne and homodyne spectroscopy) or of its high brightness (intensity-fluctuation spectroscopy).

### 2.2. *Heterodyne and homodyne detection*

In practice at radio frequencies heterodyne detection is normally used, a local oscillator being employed to beat with the incoming signal and shift its frequency to a lower one before processing. This principle is also applicable in the optical-frequency region although technically somewhat more difficult. The local oscillator signal is taken from a coherent laser source and is mixed directly on the (square-law) detector with the incident radiation field. If a photomultiplier tube is used as a detector then intermediate frequencies in the region between a few hertz and approximately $10^8$ Hz will be available for processing at the output. There is no reason in principle why any *fixed* intermediate frequency should not be chosen and the local oscillator frequency swept in order to achieve tuning, as in a normal superheterodyne radio receiver. We should then have a high-resolution device for any type of optical spectroscopy. At present, however, dye lasers and tunable lasers are not developed to the extent required to put this scheme into practice and it is more common to fix the local oscillator frequency using a given laser line and to sweep the intermediate-frequency filter after the detector.

A second difference between radio-frequency processing and optical signal processing might be mentioned at this point. This is due to the difference in energies between the microwave or radio-frequency-wave photons and optical photons, the former have an energy which is small compared with $kT$ and the latter have an energy which is normally large compared with $kT$. The result of this is that detection processes involving emission of photons are virtually absent in the optical case and as a consequence the sum frequency, which is obtained in microwave heterodyne detection, is missing. This is an interesting but not practically very important point.

Now although we have considered an arbitrary spectrum in the incident signal nothing more ambitious than a few distinct spectral lines has been studied by these optical methods to date and, for the most part, in the remainder of this paper we will consider the case where the incident signal consists of just one spectral component with, usually, a Lorentzian spectral profile. Signals of this type with frequency widths and frequency shifts within the region of frequencies which we are considering arise mainly from laser scattering experiments. A laser beam is incident upon a medium whose dielectric constant is fluctuating owing to some physical process under study and the scattered light is mixed with some of the unscattered light to act as a local oscillator. The frequency components at the output of the detector are given by

$$\begin{aligned} I(\omega) &= \int \langle I(\tau)\, I(0)\rangle \exp(\mathrm{i}\omega\tau)\, \mathrm{d}\tau \\ &= \int [I_0^2 + 2I_0\langle I_s\rangle + \langle I_s(\tau)\, I_s(0)\rangle + I_0\langle E_s^+(\tau)\, E_s^-(0)\rangle \\ &\qquad \exp\{-\mathrm{i}(\omega-\omega_0)\,\tau\} + \text{c.c.}] \exp(\mathrm{i}\omega\tau)\, \mathrm{d}\tau \end{aligned} \tag{1}$$

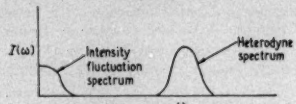



Figure 1. The spectrum of the detector output showing the intensity-fluctuation and heterodyne (Doppler) components.

and are illustrated in figure 1. The subscript s denotes signal and the subscript c denotes the coherent reference beam. $E$ denotes the electric field, the superscripts + and − denote the positive and negative frequency components and $I$ is the modulus squared of the positive frequency part of $E$ (i.e. the envelope of the field). The incident field has been assumed to be statistically stationary. The spectrum may be divided into three parts: there is the dc component due to the local oscillator, there is a low-frequency component known as the intensity fluctuation spectrum and there is the normal heterodyne or Doppler component of the spectrum. This latter will be centred at some finite frequency, different from zero, if there is any bulk motion of the scatterers giving rise to a Doppler shift. In a number of optical experiments, however, this component of the spectrum is also centred at zero frequency and we then have the homodyne case, where the local oscillator frequency is equal to the centre frequency of the signal.† In this case, if a large local oscillator power is used, the homodyne spectrum will dominate the intensity-fluctuation component, which will not be observed. At very low light levels the whole spectrum will be superposed upon a constant 'photon-noise' background.

### 2.3. *The intensity-fluctuation spectrum*

It should be noted that the intensity-fluctuation spectrum (IFS) is present even in the absence of the local oscillator. It arises from the beating together of the various parts of the signal, and in a scattering experiment gives information about the relative velocities of pairs of scattering particles. This component has been used in radar, particularly in meteorological studies, and was originally proposed for use in the identification of aircraft by radar means, the theory being that the relative motions of different parts of an aircraft would give a characteristic intensity fluctuation spectrum. In addition to being present in the absence of a local oscillator reference signal, this component of the spectrum is insensitive to phase modulation of the laser beam (the brightness rather than the coherence of the laser is used here). For this reason (bearing in mind the imperfect nature of laser oscillators as stable-frequency sources, particularly when high output powers are required) the IFS has been studied to a far greater extent in the optical region than, perhaps, in the microwave region of the spectrum. Although some of the first experiments in laser spectroscopy were homodyne or heterodyne experiments, illustrated for example in figures 2 and 3, wavefront matching and laser frequency stability problems occur and a large amount of published work since then has been concerned with analysis of the more easily obtained intensity fluctuation spectra. An important point to bear in mind with regard to this situation is that the relatively high energy of the optical photon means that, with care, optical detection is virtually noiseless in the visible region of the spectrum. The only noise is in the photon nature of the signal itself and therefore the enormous superiority of heterodyne methods in the microwave region, which arise from the ability of such techniques to dominate system noise, has not influenced the optical case. This situation has changed as the $CO_2$ laser has come into use since 10 $\mu$m detectors are intrinsically noisy even when cooled and it is fortunate that highly stable $CO_2$ lasers of relatively high powers (about 10 W) ease the problems of heterodyning at this frequency. We shall show a little later,

† Some confusion exists in nomenclature, in some optical laboratories the IFS is called a homodyne spectrum.

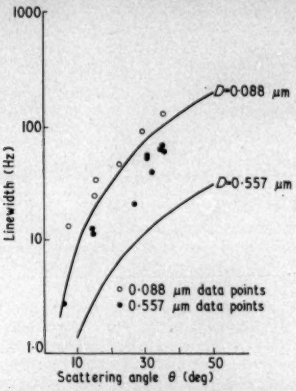



Figure 2. The results of an early optical heterodyne experiment on the Brownian motion of a suspension of PVC spheres (after Cummins *et al.*, 1964).

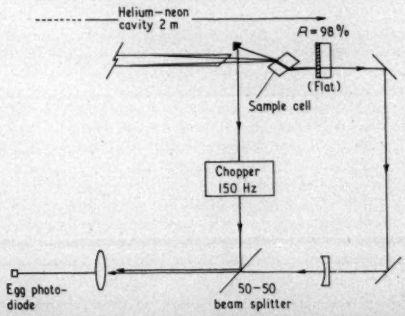



Figure 3. An early heterodyne experimental arrangement in which a 30 MHz Brillouin shift in toluene was measured (after Lastovka and Benedek 1966).

however, that owing to the statistical nature of the optical signal there is, even in the absence of system noise, a gain of somewhat over an order of magnitude to be had in the time required to do a given experiment by using homodyne rather than intensity fluctuation spectroscopy. We should note also that the optical spectrum of the incident light can only be recovered from the intensity fluctuation spectrum recorded by the detector in the case where it is symmetric and has a Gaussian amplitude distribution. For arbitrary light fields the intensity fluctuation spectrum does not uniquely determine the optical spectrum (see, for instance, Pike 1969). Many laser scattering experiments, however, give rise to light of this type and in figure 4 we show an early result using intensity fluctuation spectroscopy in which thermal transport properties of $SF_6$ near its critical point were being studied.

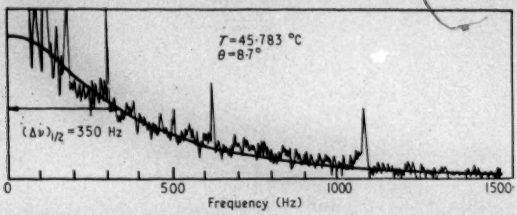



Figure 4. An early intensity fluctuation spectrum (self-beat spectrum) of light scattered by entropy fluctuations near the critical point in $SF_6$ (after Lastovka and Benedek 1966).

### 3. Signal processing

In equation (1) $I$ presented the spectral information in the form of Fourier transforms of correlation functions of the scattered field. The spectral information desired can be obtained equally well by studying the temporal correlation of the signals as by harmonic analysis. In table 1 a scheme of possible experimental arrangements is shown. Although in many experiments serial acquisition of data is employed, it is clear that a large saving in

**Table 1. Experimental techniques**

| Type \ Domain | Time | Frequency |
|---|---|---|
| Single channel | Delayed coincidence | Wave analyser |
| Parallel channels | Autocorrelator | Bank of filters |

time can be achieved by using one of the parallel acquisition schemes although when only one Doppler-shifted line is present a single tracking filter is used for following its frequency. Of the parallel processors the correlator is more flexible and easier to use than the bank of filters. A straight autocorrelator, working in real time, however, can be a costly instrument and here well established microwave techniques can come to our aid. It is well known that the spectral information contained in a Gaussian signal resides mainly in its zero crossing (van Vleck 1943). Use is made of this fact in microwave radar by hard-limiting or 'clipping' a signal before processing in such a way that one bit only is required to represent the signal as a function of time. Although there are a number of technical differences between the microwave and optical cases, which make the direct transposition of such clipping schemes impractical, the basic idea can be developed for the optical case (Jakeman and Pike 1969) with the result that a much simpler correlator can be used in optical spectroscopy. The photon nature of the detection process already gives us the information in the digital form of a random pulse-code modulated signal which is very convenient for input to fast digital circuitry. The number of photons in a sample is set to unity when it exceeds the clipping level and to zero otherwise. The normalized autocorrelation function $g^{(2)}(\tau)$ of the (stationary Gaussian) intensity fluctuations is then given by the following formulae (Jakeman and Pike 1969).

($a$) Clipping at zero photons per sample in both channels (signal and delayed signal)

$$g_{00}{}^{(2)}(\tau)=\frac{(1+\bar{n})\,2\bar{n}+(1-\bar{n}^2)\,g^{(2)}(\tau)}{1+2\bar{n}+2\bar{n}^2-\bar{n}^2 g^{(2)}(\tau)} \tag{2}$$

($b$) Clipping at $q$ photons in one channel only

$$g_q{}^{(2)}(\tau)=\frac{\bar{n}-q+(1+q)\,g^{(2)}(\tau).}{1+\bar{n}} \tag{3}$$

Such a clipping correlator is shown in figure 5 (Foord *et al.* 1969, Pike 1970, Jakeman 1970). In figure 6 we show an IFS recorded in this way; the spectrum was a single Lorentzian arising from scattering from the macromolecule haemocyanin undergoing Brownian motion; the Fourier transform of this spectrum is a simple exponential. This particular experiment was completed in a total time of 30 seconds, and the accuracy of the diffusion coefficient measured was 1%.

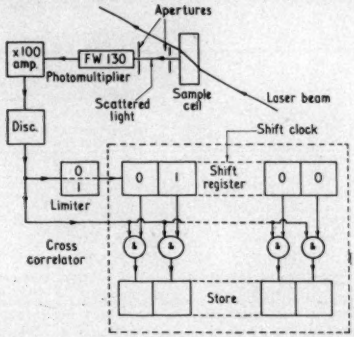



Figure 5. Block diagram of single-channel clipping correlator (after Pike 1969).

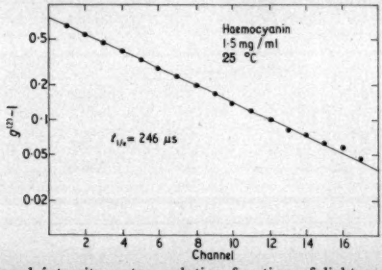



Figure 6. Single clipped intensity autocorrelation function of light scattered by the protein haemocyanin undergoing Brownian motion in solution (after Foord *et al.* 1970).

## 4. Statistical accuracy

In the last year we have made very thorough studies of the theory of digital autocorrelation, in particular we have derived expressions for the percentage accuracy of a measurement of a linewidth by this technique as a function of the various parameters involved, such as the experimental time, the number of photons per second arriving at the detector, the spectral linewidth, the number of channels available in the correlator and so on. The main results are shown in figures 7 and 8 (Jakeman *et al.* 1970 b). They show basically that for a given sample time to inverse bandwidth ratio there is a point at which no significant improvement is gained by increasing the photon flux. For instance, for a value of one tenth for this ratio this point is reached at a photon flux of approximately ten photons per coherence time. The calculations assume that only a small fraction of a coherence area is used,

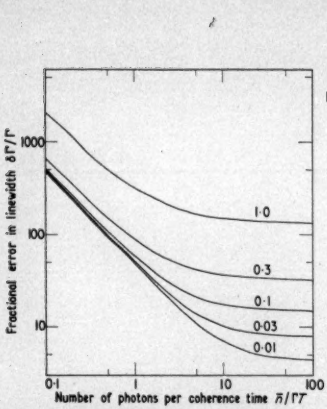



Figure 7. Percentage error in the observed linewidth of a single Lorentzian line by autocorrelation of photon counting fluctuations as a function of mean counts per coherence time, for various values of sample time to coherence time ratio.

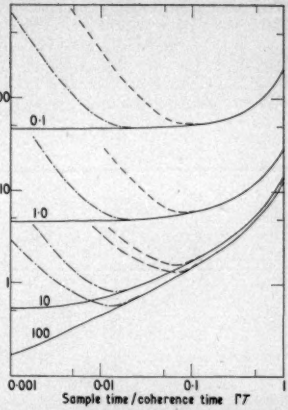



Figure 8. Percentage error in the observed linewidth of a single Lorentzian line by autocorrelation of intensity fluctuations as a function of sample time to coherence time ratio, for various values of the number of detected photons per coherence time.

the loss of signal due to the use of too large an aperture is discussed for the IFS case by Searl (1968) and Jakeman *et al.* (1970a). We have also assumed no clipping to arrive at these results so that they will not be valid if a high mean is used in a clipping correlator. The loss of correlation when clipping in one channel at zero owing to increasing mean rate and increasing sample time (Jakeman 1970) is shown in figure 9. The points are experimental values recently obtained in this laboratory.

The reduction of the error in the measurement of the correlation function by the use of a homodyne reference beam is given by the expression

$$\frac{\mathrm{Var}\, g^{(2)}(\tau)_{\mathrm{IFS}}/(g^{(2)}(\tau)_{\mathrm{IFS}}-1)}{\mathrm{Var}\, g^{(2)}(\tau)_{\mathrm{Hom}}/(g^{(2)}(\tau)_{\mathrm{Hom}}-1)} = 8\left(1+\frac{1}{|g^{(1)}(\tau)^2|}\right) \tag{4}$$

$g^{(1)}(\tau)$ is the autocorrelation function of the optical field. This improvement is shown in the experimental results shown in figure 10. The accuracy of linewidth measurement with a fixed time of experiment is improved by a factor of approximately four in the homodyne case; this would correspond, alternatively, to a reduction in experimental time for the same accuracy by a factor of sixteen. The change of clipping level from zero to one shows little difference in this case.

## 5. Applications

A series of results obtained recently in various applications of this type of optical spectroscopy will now be discussed. Systems capable of study by optical means which show fluctuations in the frequency range up to 100 MHz include the motion of macroscopic objects, the motion of macroscopic molecules and the motion of microscopic excitations in solids, liquids and gases.

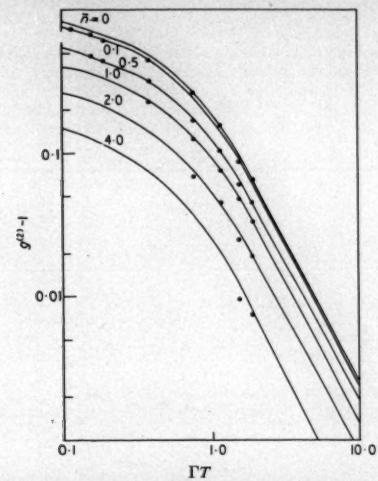



Figure 9. The loss of correlation due to integration with single-channel clipping at zero. The mean photon number per sample time is shown against the curves. The dots are experimental points.

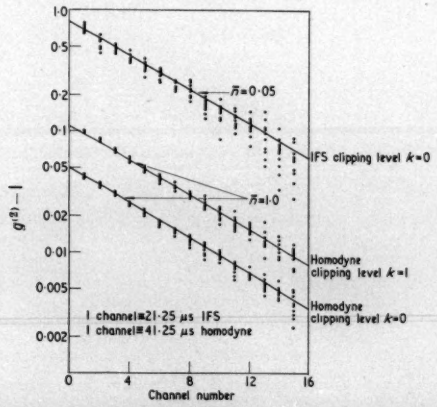



Figure 10. Homodyne and intensity fluctuation spectra of scattering from haemocyanin undergoing Brownian motion. The different spreads in experimental points show the gain to be achieved by adding a reference beam as local oscillator. The scattered intensity was the same in the three cases.

5.1. *Anemometry*

In the first category we include the general topic of velocity metering and anemometry where bulk motion is studied or where flow and turbulence are studied by scattering from suspended indicator particles. A typical apparatus for studying flow and turbulence in pipes is shown in figure 11 and some results obtained with this apparatus are shown in figure 12 (Bourke *et al.* 1969). The same principles can be applied to the observation of atmospheric wind velocity and turbulence and in figure 13 some results obtained (Hughes

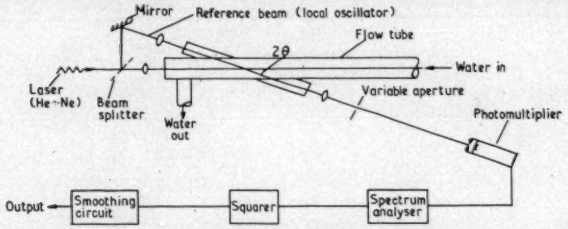



Figure 11. An experimental arrangement for measuring flow and turbulence in pipes (after Bourke *et al.*, 1969).

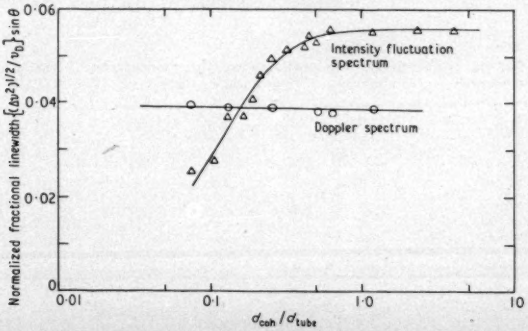



Figure 12. Experimental results obtained with the apparatus of figure 11 showing the information gained about the spatial structure of turbulent flow by using the IFS technique (after Bourke *et al.* 1969).

*et al.* 1970 a) using an argon laser at a range of 100 ft are shown. More recently the $CO_2$ laser shows promise for this application and in figure 14 the result of a laboratory experiment at 10·6 $\mu$m using a cooled lead–tin telluride detector are shown. In figure 15 is shown a schematic arrangement for obtaining three vector components of wind velocity taken from recent work by Huffaker *et al.* (1970) in which they have successfully measured wind velocity at short ranges in one channel of such a system.

5.2. *Macromolecules*

Study of the Brownian motion of macromolecules has been a particularly successful application of the methods of homodyne and intensity fluctuation spectroscopy (Cummins

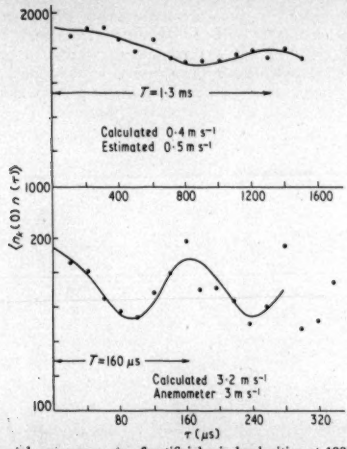



Figure 13. Experimental measurements of artificial wind velocities at 100 ft range using an $Ar^+$ laser and digital autocorrelation (after Hughes *et al.* 1970).

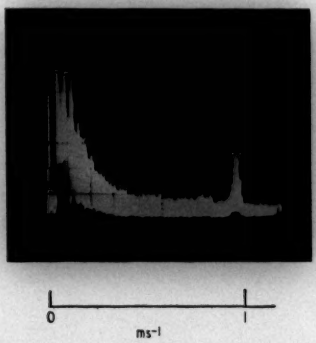



Figure 14. An experimental laboratory measurement of wind velocity using a $CO_2$ laser and wave analyser.

*et al.* 1964, Dubin *et al.* 1967). Recently the new autocorrelation methods have made it possible to study very small concentrations of proteins and to obtain results in short experimental times. In figure 16 we show the concentration dependence of the diffusion constants of three globular proteins down to very small concentrations with high experimental accuracy

(Foord *et al.* 1970). In figure 17 we show how the technique can be used to study a mixture of two different macromolecules, in this case a mixture of haemocyanin and alpha-crystallin, and again find the diffusion coefficients to good accuracy. Fractionated polymers can also be studied by these methods and in figure 18 is shown an experimental result on polystyrene in toluene. Polymer solutions tend to scatter less light than proteins and in this example the

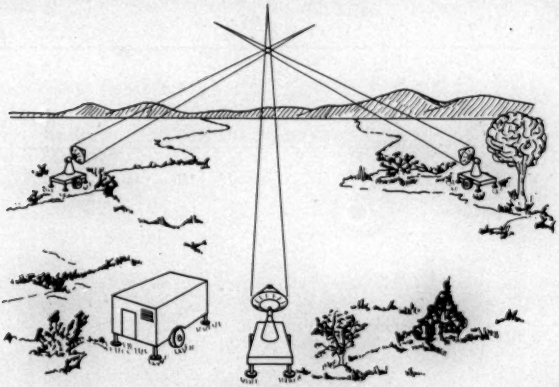


Figure 15. An arrangement for the measurement of three components of wind velocity (after Huffaker *et al.* 1970).

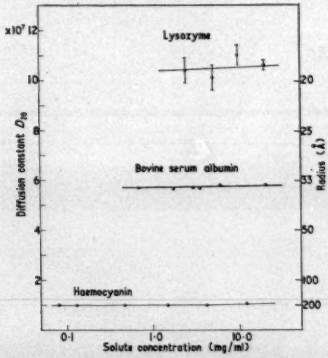



Figure 16. The concentration dependence of the translational diffusion constant of three protein molocules in solution by a single clipping digital correlator IFS measurement.

scattering rate into the detector was $6 \times 10^{-3}$ counts per sample. The extra accuracy obtained by homodyning is here put to good advantage. These last two results are from

work, as yet unpublished, in my own laboratory. Another example of this type in which more than one spectral component has been studied is the work of Cummins *et al.* (1969) on the rod-like tobacco mosaic virus. They were able, using a wave analyser IFS technique to separate bulk diffusive motion from rotational diffusion by observing the angular dependence of the spectra recorded. An increasing number of publications are appearing each year both on the theory and on experiments on light scattering from macromolecules and it seems that the method is now firmly established as a tool in a number of laboratories.

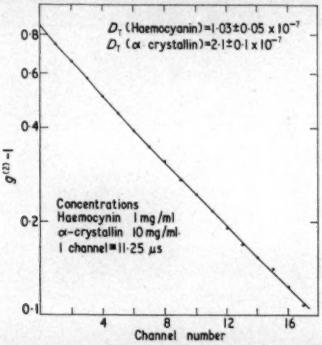



Figure 17. The observation of translational diffusion constants in a mixture of two proteins in solution by a single-clipping, digital correlator IFS measurement

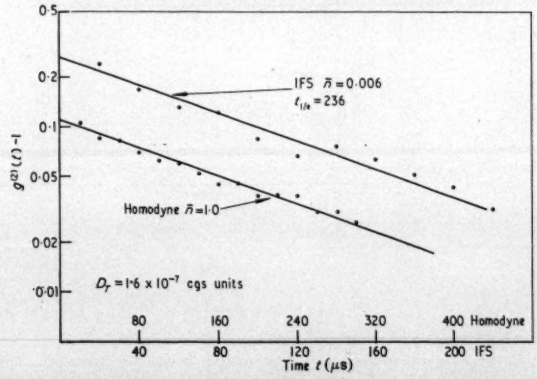



Figure 18. Measurement of the translational diffusion constant of a fractionated sample of polystyrene in toluene by single-clipping, digital correlator IFS and homodyne techniques.

### 5.3. *Microscopic excitations*

In the category of the study of microscopic motions we show first an experiment (Hughes *et al.* 1970 b) in which the strain field due to a stimulated phonon flux in an acoustoelectric

oscillator is studied by a heterodyne method. The apparatus is shown in figure 19, and in figure 20 are shown the resulting Brillouin spectra at different positions across the slice of the oscillator. This shows clearly that different acoustic modes are effective in different regions of the oscillator and provides a technique for spatial study of phonon flux which would be extremely difficult to perform by any other means. Another recent example of the study of microscopic motion is given in figure 21 where we show the intensity fluctuation spectrum, this time taken using a wave analyser, of the Rayleigh light scattered by entropy fluctuations in a pure liquid, namely benzine. This is a recent version of the experiment illustrated in figure 3 and gives a measurement of the thermal diffusivity of the liquid.

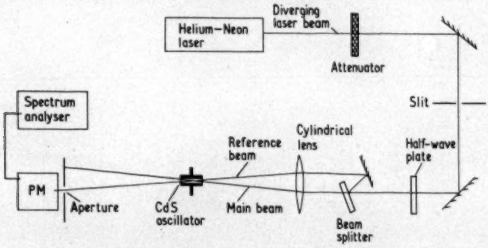



Figure 19. An experimental arrangement for the measurement of acoustoelectric strain fields in a CdS oscillator by heterodyne detection (after Hughes *et al.* 1970).

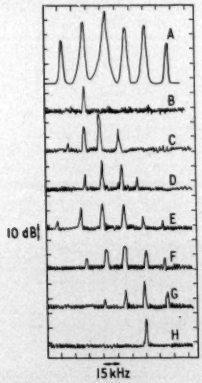



Figure 20. Results obtained with the apparatus of figure 19 showing the different phonons excited in different regions of the oscillator.

Transport properties near the critical point of simple fluids and of binary mixtures have been widely studied by IFS techniques. In figure 4 we showed an early example of this type of study. Curiously $SF_6$ is still a question of some controversy (Braun *et al.* 1970, to be published) but more recently xenon has been carefully studied (Henry *et al.* 1970) and

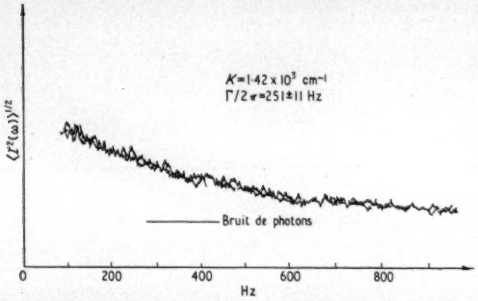



Figure 21. A measurement of the decay rate of entropy fluctuations at constant pressure in benzine, giving the thermal diffusivity by a wave analyser IFS technique (after Bergé and Dubois 1969).

the results have been used to dispute the validity of proposed relations between critical indices. The critical behaviour of binary mixtures gives similar data for theoretical discussion (Bergé *et al.* 1970, Chen and Polonsky-Ostrowski 1969, Pusey and Goldberg 1969, Chu *et al.* 1968).

These examples illustrate a wide range of phenomena which can now be studied by optical spectroscopy. To compare with classical optics we could point out that the Fabry–Perot interferometer or the Fourier multiplex interferometer barely come on to the lower end of the scale of resolution which we have been discussing and although up to the present the spectra which have been studied have been relatively simple there is no reason to suppose that as the subject develops more and more complex phenomena will not receive attention.

## References


Benedek, G. B., 1968, *Polarisation, Matière et Rayonnenent* (Paris: Presses Universitaire de France).

Bergé, P., Calmettes, P., Laj, C., Tournarie, M., and Volochine, B., 1970, *Phys. Rev. Lett.*, **24**, 1223–5.

Bergé, P., and Dubois, M., 1969, *C.r. Acad. Sci. Paris*, **269**, 842–5.

Bourke, P. J., *et al.*, 1969, *Phys. Lett.*, **28A**, 692–3.

Braun, P., Hammer, D., Tscharnuter, W., and Weinzierl, P., 1970, to be published.

Chen, H., and Polonsky-Ostrowsky, N., 1969, *Optics Commun.*, **1**, 64–6.

Chu, B., Schoenes, F. J., and Kao, W. P., 1968, *J. Am. Chem. Soc.*, **90**, 3042–8.

Cummins, H. Z., Carlson, F. D., Herbert, T. J., and Woods, G., 1969, *Biophys. J.*, **9**, 518–46.

Cummins, H. Z., Knable, N., and Yeh, Y., 1964, *Phys. Rev. Lett.*, **12**, 150–3.

Cummins, H. Z., and Swinney, H. L., 1970, *Prog. Optics*, Vol. 8, Ed. E. Wolf (Amsterdam: North Holland).

Dubin, S. B., Lunacek, J. H., and Benedek, G. B., 1967, *Proc. N.A.S.*, **57**, 1164–71.

Foord, R., Jakeman, E., Jones, R., Oliver, C. J., and Pike, E. R., 1969, *Proc. IRE, Conf. Proc.*, No. 14, p. 271.

Foord, R., Jakeman, E., Oliver, C. J., Pike, E. R., Blagrove, R. J., Wood, E., and Peacocke, A. R., 1970, *Nature, Lond.*, **227**, 242–5.

Henry, D. L., Swinney, H. L., and Cummins, H. Z., 1970, *Phys. Rev. Lett.*, to be published.

Huffaker, R. M., Jelalian, A. V., and Thomson, J. A. L., 1970, *Proc. IEEE*, **58**, 322–6.

Hughes, A. J., Oliver, C. J., Pike, E. R., and O'Shaughnessy, J. J. B., 1970 a, *RRE Res. Rev.*, **9**, 7.

Hughes, A. J., Maines, D., and O'Shaughnessy, J. J. B., 1970 b, *J. Phys.* D: *Appl. Phys.*, **3**, 751–7.

Jakeman, E., 1970, *J. Phys.* A: *Gen. Phys.*, **3**, 201–15.

Jakeman, E., Oliver, C. J., and Pike, E. R., 1970 a, *J. Phys.* A: *Gen. Phys.*, **3**, L45–8.

JAKEMAN, E., and PIKE, E. R., 1969, *J. Phys.* A: *Gen. Phys.*, **2**, 411–2.
JAKEMAN, E., PIKE, E. R., and SWAIN, S., 1970 b, *J. Phys.* A: *Gen. Phys.*, **3**, L55–9.
LASTOVKA, J. B., and BENEDEK, G. B., 1966, *Physics of Quantum Electronics*, Eds P. Kelley, B. Lax, and P. E. Tannenwald (New York: McGraw Hill), pp. 231–40.
PIKE, E. R., 1969, *Riv. Nuovo Cimento*, Ser. 1, **1**, Numero Speciale, 277–314.
PUSEY, P. N., and GOLDBERG, W. I., 1969, *Phys. Rev. Lett.*, **23**, 67–70.
SEARL, O. B., 1968, *Phys. Rev.*, **175**, 1661–8.
VAN VLECK, J. H., 1943, *Harvard University Radio Res. Lab. Rep.* No. 51.

C. L. BOLTZ

# Information: the essential commodity

## Thoughts based on the Paris seminar of March 1970

'Pollution by documentation' was one of the expressions coined by the Rt Hon. Anthony Wedgwood Benn, then Minister of Technology,† at a seminar arranged by the Organization for Economic Co-operation and Development (OECD) near Paris last spring. He meant, as he said, the wrong document in the wrong place at the wrong time, an apt description of one aspect of the information problem. Every physicist in industry is aware of the mass of documents now being produced, whether quoted as tons weight or numbers of documents; it is a familiar topic. So is the regretful comment from many a physicist that he does not get time to read everything he should. Even the *Physics Bulletin*, probably the most easily readable of all physics journals, is not thoroughly read, as many members of the IPPS have confessed.

The information business itself has become moreover a specialization. So conferences on the subject tend to be filled by people who discuss the subject among themselves as a special technology, with little of the discussion getting to the people who need to use information, the working scientists. Several speakers at the 4-day seminar (at which there were representatives from 19 countries) noted indeed that less than a third of the hundred participants were from industry, though the subject of the seminar was: *Government Responsibilities in Information for Industry*. It was an official gathering of representatives who were to return to their countries with recommendations about government action. There was considerable cut-and-thrust between advocates of more government action and those who believed in private enterprise. However, this is not a report on the seminar; my purpose is to extract from the proceedings much that emerged, even if incidentally, of interest to physicists, some of whom may be interested as users and providers of information and some of whom may be interested in the more general problem of dealing with the information explosion.

One outstanding contribution came from Professor T. J. Allen, Assistant Professor of Organizational Psychology and Management at the Sloan School of Management at MIT. It summarized six years of study of the flow of information in research-and-development projects in many firms. Some physicists may squirm at the word 'psychology' so it is important to realize that American psychology (as distinct from psychiatry) is behaviourist and is thus broadly committed to the thesis that the only measurable evidence of the mechanism of the human mind is in human behaviour. Much of the experimentation in this field is in fact very akin to physics and it is worth recalling that in Dr Broadbent's team at the Applied Psychology Unit in Cambridge there are physicists and engineers. The work of Professor Allen and his colleagues, which was supported by money from the National Science Foundation, was of this sort.

The first set of experiments led to contradictory conclusions. The team studied 21 R and D projects involving more than 250 engineers and scientists working on separate problems. Each man (or woman) was studied in real time while at work, the aim being to find out whence the researchers obtained their ideas. There are in essence only three types of information available in any organization. First there is documentary information, the scientific and technical literature, which must obviously be generated for the most part

† The Ministry of Technology has been replaced by the Department of Trade and Industry but so far few details have been given of the difference of approach suggested by the Secretary of State. Whenever the Ministry of Technology is mentioned in this article, therefore, the reader can substitute DOTI without invalidating the argument. Occasionally 'Ministry of Technology' is correct in historical context.

outside the organization. Secondly there are the personal sources of information, some of whom would be inside the organization and, thirdly, some outside. The Allen team broke down sources into smaller parts—pieces of ideas so to speak—because it was obvious that no conceptual idea would come complete from a single source. When the sources were at length analysed it was found that the scientists and engineers obtained only 11% of their ideas from written material. This was the first important conclusion. This meant that 89% of the ideas must have come from *people*. The use of documentary information by the engineers and scientists was thus very low.

It should be noted that the projects studied were high-level ones concerned mainly with aerospace research. The researchers were, as Professor Allen put it, 'top-flight' engineers and scientists working on systems that might be seen as hardware in from 10 to 20 years' time. They were not routine men and women working on a standard selling project and its improvement. In other words they were people to whom new ideas were very important.

Having established that the usage of literature was low, the team examined the impact of oral information from *outside* sources and found that there was a negative correlation between a man's performance inside and the extent to which he relied on people outside the organization for information. On the other hand there were high positive correlations between performance and oral sources *inside* the organization. These results have been confirmed by a number of other investigators, and the Allen team did a study of more than 200 different organizations at this project-team level with results substantially the same as those mentioned.

It was obvious that information was getting into the organization somehow, for it was a safe assumption that no organization could be self-sustaining, especially in such advanced technology. Yet the literature was not used very much by the bulk of the researchers and the oral information from outside was not effective. This set Professor Allen a dilemma. As he said, he and his colleagues were assuming that 'every engineer in the organization was just like every other engineer in the organization with respect to his information-seeking behaviour and his role in the communication network of the organization'. This was found not to be true at all.

This failure led to Professor Allen's later significant discoveries. It has been described here as an exercise in better communication to the reader. Had the work been written in the style of a scientific 'paper' it would have been less effective, for as Sir Peter Medawar wrote (in his essay on hypothesis and imagination in his book *The Art of the Soluble*), papers 'not merely conceal but actively misrepresent the reasoning that goes into the work they describe'.

Professor Allen saw that some rethinking was necessary. At length he decided to change his research tactics. He stopped studying projects and began to study organizations and the structure of their communication methods. He and his colleagues looked at eight different organizations in three different countries, asking the question: who communicates with whom within the organization? The routine technique was to ask questions of everybody in the organization. He then drew a diagram showing arrows entering and leaving any individual, indicating communication transfers. He then selected all individuals with one standard deviation above the mean number of arrows entering an individual 'node' and examined them separately. He found that these selected people read the professional engineering and scientific journals far more than others. The result was the same for a small laboratory of 28 professional scientists and engineers and for a large aerospace firm of 341 people. Other types of literature, however, were read to about the same extent by the specially-selected people and others. Furthermore it was found that these special people maintained outside contacts much more. 'The average member of the organization did not read the professional engineering and scientific journals but some people in the organization *did* and then the information came from that literature to this small set of people within the organization and through them to the average engineer in the organization.' 'Information comes best from people outside the organization to this small set of people who maintain outgoing relations with them within the organization and then

from them to the more average member of the organization.' 'In other words it is a two-step process involving an intermediary who maintains ongoing contact with these two types of information-source and performs a translation process and puts the information that is contained for example in the professional literature into words that can be understood by the average engineer.'

This was the major discovery, the existence of the special communicators. He dubbed them 'technological gatekeepers' and this concept is becoming part of the accepted body of thought in the information sciences. The term is awkward but it is difficult to coin a better one. English people might prefer 'watchdog'. One British government representative suggested 'traffic cop' as more representative of the function. None of these has the stamp of an important technical term. A possibility exists in TIE for Technical Information Exponent, where the acronym hints at the function, for these individuals do connect the outside with the inside.

What sort of people are these gatekeepers? They are not merely communicators. They are first-class scientists or engineers themselves and they hold more patents and publish more papers and receive higher ratings by management for their technical performance, according to Professor Allen. They are also usually supervisors with six or more scientists reporting to them.

This description of the gatekeepers raises implications. Their quality means that they get promotion, so that they move into the administration and are lost as gatekeepers. Will there be others to take their places and will the new ones need some time to become effective? If so this could mean that an organization would go through a poor period of less effective communication. Has this happened? A further implication is that managements will in time be forced to reconsider the structure of their organizations, not just to improve the flow of information but for another reason. At present, as Dr Peter expressed entertainingly in the Peter Principle, people tend to be promoted until they reach their level of incompetence, and in due course the undoubted truth behind the wisecrack will have to be faced and there will have to be some way of raising the status of men doing a task supremely well without promoting them away from their jobs. (There are some organizations already where special awards can be made, carrying higher salaries, without taking men out of their most effective spheres, though this is done, one suspects, because the higher positions are already full rather than to keep men to their best speciality.)

Another question is: Are they born or made? Professor Allen thinks that they are made, not born. Education helps and qualification as a PhD gets them into habits of searching literature and using libraries. However, this is not the whole truth, for more information is gained and transmitted orally than is culled from literature. This truth has been stated by many practitioners and Mr Benn has mentioned that this was one of the empirical findings of the Ministry of Technology. Allen further thinks that the emergence of gatekeepers is more likely due to the way people are treated inside an organization. So his team is now following people from the time they enter an organization to discover which ones become gatekeepers and which do not. He does not believe that the gatekeeper function can be formalized, as some people have suggested. Furthermore the information is given both formally and informally, and gained in both ways as well, and so any rigidity would be harmful.

He has found that a similar system operates in countries, especially small countries. His team has applied the same sort of investigation and analysis to a small country as a pilot study and found again that there are these stars of communication. They are the ones who attend more scientific and professional conferences in a country than do the average members and, moreover, attend more such conferences in foreign countries and report more communication with foreign countries. '*I think*,' he has said, '*that it is a fairly universal phenomenon that we are uncovering*'.

One of the analytical techniques used by the group came from graph theory. Networks showing communication links were drawn and from these, by applying graph theory, 'strong' components could be extracted. When this was done it was found that the gatekeepers were among the strong components. Furthermore, these strong components did not line up

with the structure of the organization being investigated. The gatekeepers were found to have their own tightly-connected network, and were in close contact with one another. The average members of the organization communicated with the outside through a whole network, thus increasing the probability of any member getting what he wanted in ideas and information.

If one accepts the findings from this continuing and attractive piece of operational research by Professor Allen—and as far as I know nobody has advanced any logical arguments against his methods or disagreement with his conclusions—there are clearly some further inferences to be drawn. First there is the clear message for management and government that many organizations need to be restructured to take full advantage of the gatekeeper-network concept. It is interesting to recall in this connection that Mr Benn himself said that his Ministry had discovered that the management problem was one of structure and without the right structure we could not hope to establish any of the things that really determined whether we could generate and use information. It is not suggested here that he was thinking of the gatekeeper-network or indeed knew about it, yet nevertheless the discovering that restructure was necessary to help information processes does bear out the more explicit inferences from the Allen research.

The importance of verbal information has already been mentioned. It leads to the further implication that conferences should be organized to give much more scope to informal discussions between members than they usually are. This is not entirely a new idea, for many physicists have stated the importance of the informal discussions, but the fact that this notion is a result of investigation instead of being merely a sum of personal opinions is important.

Informality comes into the structural modification needed to assist the flow of information. The task of a research director is seen to include that of spotting who are his communicators, his gatekeepers, and arranging things to make their activities easier. This could involve, as Professor Allen has pointed out, a look at the sheer mechanics of communication such as the layout and architecture of a building. His group looked at people in organizations and measured the distance from one man's desk to everybody else. They went outwards in concentric circles and looked at the proportion of people with whom each person communicated as a function of the distance of separation. The result, not surprisingly, was that the probability of communication was a function of this distance. (As a first effort it was taken as inversely as the square of the distance, which in one investigation led to the conclusion that at 25 metres distance the probability of communication was down to the asymptote, i.e. almost nothing!) His group has now been engaged as consultants in the reorganization of organizations and even in the design of a new building, the occupants of which have now moved in. So it is hoped that experimental results will be available next year. One corollary to this is that the much-favoured modern design of laboratory with long corridors off which small offices and laboratories are arranged serially would seem to be very poor for optimum communication. One is left with the tantalizing vision of perhaps a spiral plan or concentric plan, with the gatekeepers readily reachable.

Another conclusion from all this work, which has proved so germinal, is that the task of outsiders to communicate with a firm is made easier because what they have to do is go to the information gatekeepers (who are not the public relations officers). Firms that distribute handouts on new products could pay attention to this. Far more importantly high-level publications must go to them. So the first task of journal publishers is to find out the gatekeepers. Thus the number of people to be kept informed in an organization is reduced and the task simplified if one can identify the gatekeepers. How this is to be done remains a challenge.

This is about as much as we can construe from Professor Allen's work. Dr Alexander King, Director General for Scientific Affairs in OECD, said at the beginning of the conference: 'We know far too little about how information actually flows within industry. One thing that is clear is that we should be talking certainly not of documentation in any narrow sense but industry in a deep and very comprehensive sense in which the spoken word, as has

already been mentioned, is an integral and perhaps one of the most important parts.' Professor Allen's summary account of his investigations did a lot to show in understandable terms how to fill in the gaps in knowledge mentioned by Dr King. It also demonstrated that fairly simple experimentation on behavioural lines could yield usable knowledge in the information field. Professor Allen's work, however, was restricted to the mechanism of the flow of effective scientific and technical information within an industrial organization and, by extension, within a country, and though his findings had implications already discussed there still remain many other aspects of communication of interest to the practising physicist.

Though we may think we can leave large firms to look after their own information systems and hope that they will learn from investigations such as those of Professor Allen, who has said himself that his findings will lead to drastic changes, we are still faced with the problem of the smaller firms, and this problem exists in a number of countries as well as Britain. It has been argued, and it was repeatedly contended by the chief apostle of private enterprise at the seminar, Mr W. Knox of McGraw Hill, in what Mr Brady of the National Bureau of Standards pilloried as the 'gospel according to St Knox', that businesses that show no enterprise or innovation should be allowed to fail. As Mr Knox put it, the privilege of private enterprise is not only to make a profit but also to fail. Is it right to put taxpayers' money into small enterprises that would never put their own money or effort into innovation? Part of the reply is that small businesses are so much the lifeblood of Britain and some other countries that it would be a national disaster to let a large proportion of them fail. According to data presented graphically to the seminar by Mr Cyril Coffin of the Ministry of Technology there are some 53 000 businesses in Britain employing less than 25 people, some 27 000 employing between 25 and 200, about 5000 with staffs of between 200 and 500, and 3000 employing over 500. Thus some 90% of all businesses employ less than 200 people, and they contribute altogether about 30% of the total annual value in sales. The figures can be tabulated thus:

| Number of employees | 1–24 | 25–199 | 200–499 | Over 500 |
|---|---|---|---|---|
| Number of firms | 53 000 | 27 000 | 5000 | 3000 |
| Annual sales | £2000m | £6000m | £5500m | £13 000m |
| †Annual sales per firm | £40 000 | £220 000 | £1m | £4m |
| †Average number of employees | 12 | 112 | 350 | ? |
| †Annual sales per employee | £3300 | £2000 | £3000 | ? |

† These figures are my rounded arithmetic and must not be attributed to Mr Coffin. One striking fact is that the very smallest firms have a higher productivity than any others, perhaps an indication of the dead weight of administrative and service personnel in the larger firms.

One of the difficulties with an information service is that there are many small businesses that do not see any problem and never ask for information. Should one let them fade out? The French government is concerned with this and there are services in Denmark, Italy and Japan catering for the small business. The USA has its Small Business Administration but its State Technical Services Organization is being phased out. In Britain there are several services (the whole activity of the governmental involvement in information for industry, including the existence of OSTI, is considerable and would merit a long article) and of special interest to the small business is the system of industrial liaison officers (ILO's). Many physicists will be aware of this service and in any case the Ministry of Technology produced informative booklets when the services were established. It is therefore only necessary here to outline the system.

There are 85 ILO posts at 75 centres in Britain. Each centre has a secretary and the officers are engineers or scientists of graduate status. These officers are members of the staffs of colleges of technology and technological universities, for the siting of such institutions corresponds to industrial development and the centres are in these establishments. At the moment of writing two-thirds of the cost of running the centres is borne by the

Ministry of Technology and the rest by the educational authorities. As far as the user is concerned the ILO service is free. The average number of firms to each ILO is about 700. Some 95% of firms are within an hour's drive from an ILO centre. (On the whole it is found that it is the firm with about 300 employees that uses the service most.) Each ILO has to act as a sort of missionary preaching productivity though this role is now declining and will continue to do so. He can help to educate firms generally by means of lectures and by influencing his own educational establishment. His main job is to provide a service for firms in the form of information, diagnosis and reference to sources. With a serious enquiry he visits the firm, identifies the problem, decides what best specialist advice can help and where it can be obtained. He follows up progress. He thus himself needs a well-organized information system and is aware of what is happening in his own establishment, in research associations, other research laboratories and so on. The locating of the ILO centre in an establishment of technical education has proved helpful in that it helps to link local industry with the college, makes immediately available locally a great deal of the information needed by a firm and stimulates teachers in the college. It has been shown that in colleges where the principal and staff are enthusiastic some 45% of advice has come from them rather than from outside. So far each ILO has made about 140 visits a year, and the aim is to build up to 150.

It should be mentioned that for many years the USA's Department of Agriculture has had at least one agent in every state doing work like that of the British ILO's. This has produced an agricultural revolution.

The most striking statement made by Mr Coffin related to the cost-effectiveness of the ILO system. A centre with one ILO costs between £7000 and £8000 to run. Consultants called in have estimated benefits to industry, as a result of enquiries to ILO's, for some ten centres, randomly chosen. It comes to about £50 000 a year per centre. The benefit-to-cost ratio is thus between 6 and 7 to 1. This is admittedly subject to reservations and qualifications but even if it were reduced somewhat it would still be a high profitability. Mr Pavitt, OECD's Head of the Unit for the Economic Aspects of Science and Technology, questioned the 7-to-1 figure. No doubt thinking of the profitability he asked why private firms did not do this. His question is understandable and Mr Coffin did not really answer it. But we can make a guess at some of the answers. The ILO, as the system is run, has at his bidding a great availability of expertise from the teaching staff and from government-run or government-aided establishments. This is not to say that all the help sought by an ILO is free, though a lot of it must surely be so, but it is obvious that help as part of a national service is a very different matter from help as a means of extending an entrepreneur's profits. All the same, it looks as if an opportunity exists for private *information* consultants for small industries as distinct from private consultants who are selling their own research capacity. There do exist services such as that of the Engineering Capacity Exchange, which utilizes information on production capacity, and Derwent Publications, which sells information on patents, half of which is used by research and development laboratories, and one publicity expert, formerly with a large American firm in Britain, has set up a consultancy 'in communications' but this it would seem is concerned with sales and not with research and development. There does, therefore, appear to be an opening for well-qualified physicists who can deal in information, though of course their costing would have to be different from that of an ILO.

There might also be scope for private enterprise in the establishment of information analysis centres. Such a centre, as Mr Coffin said, consists of 'a group of scientists or technologists who have accepted the responsibility of gathering together everything known about a particular well-defined subject field, of evaluating this information, and of communicating it to others.' More than 100 such centres exist in the USA and subjects currently being considered as suitable for such centres in Britain include cryogenics, welding and electronic materials. Initial costs for such centres, which need computers for indexing and retrieval, are high—hence the case made for government support. However, there is nothing to prevent a computer agency with a powerful computer available from going into

such a business if there are enough users to make it profitable. It would have to recruit high-quality scientists who are specialists in the chosen field.

Mr Coffin mentioned the possibilities for the Small Business Centre. One has been established by the Ministry of Technology, at the University of Aston, and two others are being considered. Such a centre provides a service for the local business community. The Aston Centre is oriented towards management but the two that were being actively considered at the time he was speaking were to include technology.

The role of libraries and librarians is another aspect of the information problem. Today there is a well-trained body of professional librarians and ASLIB has been greatly concerned in raising the status of these people in the scientific and technical field. Most libraries are places of reference and modern librarians are normally anxious to help, so unlike so many of older inward-looking librarians who seemed to live in a world of their own. Nevertheless, as a whole, libraries are stores of documents and they range in size from one small room to enormous places such as the Reading Room of the British Museum. But they do remain, so to speak, largely historical, and Dr C. van Dijk, Director of the Library and Scientific Information of the Royal Netherlands Academy of Sciences, said that there is a new trend to make a library a well-functioning source of information, a tool for users. User studies, he considered, should form part of the curriculum of higher library education and special industrial librarians should be instilled with a better understanding of the industrial environment in which the library functions.

A first-class example of how a library could function is seen in the National Science Library of Canada. It is the major agency responsible for disseminating information to the scientific, technical and industrial communities of Canada. Its chief librarian, Dr Jack E. Brown, gave an account of its activities to the seminar. The National Science Library provides a question-and-answer service for scientific and technical information, a national data service, a loan service and a photocopy service. It is a referral centre for information not available in its own records. It is a clearing house for technical reports and government publications. It advises and assists industrial firms in the establishment of information centres. It conducts experiments on new information-processing techniques. It acts as a translations information centre. And it does all this on an annual budget of 1·5 million Canadian dollars and with a staff of just over 100.

He gave as an example some data on the industrial usage. The National Science Library processes between 300 and 400 photocopies or loans every day. Of the requests for this service about 40% comes from industry, which amounts to some 50 000 photocopies or loans for industry every year. A quarter of the requests for searches in literature and for the compilation of bibliographies comes from industry. About 25% of the subscribers to the national data service are industrial and Dr Brown thinks this low figure is due to the heavy weighting of the data bases towards the natural sciences, but this is changing. Recently there has been a statement of government policy according to which all information services come under the National Librarian, which means that though details may change the National Science Library is officially given a high status in the dissemination of scientific and technical information.

Dr Brown's library is a national one but the positive attitude towards the provision of information could well be extended to all libraries in industry. There is no doubt that many libraries in small establishments are excellent sources of information, with copying services immediately available and librarians who are funds of information, but it has to be asked for. There is surely a case for the re-examination of the functions of libraries to make them outgoing sources of information.

There is one ancillary point in connection with the use of documentation and it was raised at the seminar. It is the matter of classified information derived from defence research. A lot of this can be declassified and used by everybody. The USA has clearing houses for such information, and it is a part of the National Bureau of Standards in Washington. Mr Brady disclosed that some 18% of its sales in 1969 was to customers outside the USA. The clearing-house has a staff of 300 and sold three million documents

in 1969. There is also a report centre in Britain to make such information available and most physicists know how much easier it has become to visit defence establishments, where a lot of original work is done. We have seen evidence of its discoveries in such matters as carbon fibres (Royal Aircraft Establishment), lasers and neutron generators and semiconductor lamps (Services Electronics Research Laboratory), silicon nitride ceramics (Admiralty Materials Laboratory), and so on. This sort of information is scattered in journals and proceedings and there may be a case for the channelling of it into one publication for general circulation. It is at present not always easy to track down the originator of a development. The publicity on carbon fibres, for example, must have given the impression that one particular firm was solely responsible, and there are firms that commercialize under licence who do not disclose that a particular laboratory has been the originator. Such behaviour tends towards a falsification of information in the fundamental sense. It can perhaps be argued that such information is of interest only to the historian of science but surely truth is an important part of information as a commodity.

The functions and *raison d'etre* of professional, scientific and technical societies are a recurring topic of conversation in and around the corridors of research. Clearly, they provide methods of information transfer and some of them were created originally for that purpose. The most renowned of them all, the Royal Society, was formed by Wilkins and others in the seventeenth century in order to share information. In the course of time the functions of many of these societies have changed, and they also vary from country to country. In Germany, for instance, they do not provide professional status whereas in Britain many do. Every such society has from time to time to re-examine its functions and assess what it provides for subscribing members. The single exception probably is the Royal Society but most societies are not in this category and some have budgetary troubles that are frequently under review. In this present context it is their function in the information field that is of interest and here they have very special opportunities, for in many cases they are big enough and well funded so that they can do things that small organizations could not do, not only because they have money of their own and can publish in the open market in a way envied by private publishers, but also because they are sometimes in a position to persuade governments to supply money.

Dr Robert Cairns, Vice President, Research, and Director of Hercules (USA), went very thoroughly into the activities, both achieved and potential, of such societies with special reference to the American Chemical Society. Mr J. Larink, Director of Publishing for the German Engineers' Association, gave an account, a fairly pessimistic one, of what was happening in Germany. Dr Cairns's account provides an excellent example of the possibilities and the doubts. 'Why,' he asked, 'do 20 000 chemists subscribe to the journal of the American Chemical Society? What proportion reads how many of its 8000 pages a year? Is this the most effective way to spend one million dollars?' The secondary service, concerned with abstracts, is another story. *Chemical Abstracts* covers 12 000 journals throughout the world and patents issued by 25 countries. Some 280 000 abstracts were published in 1969. Each volume has seven types of indexes. Co-operative input and marketing arrangements have been achieved with a consortium headed by Britain's Chemical Society and a group led by Germany's Gesellschaft Deutscher Chemiker, and other arrangements are being negotiated. A number of computer-based information centres are being established in the world. There is also the *Chemical Compound Registry*, a computer-based file linking information on structure, names and bibliographic references on all chemical compounds. The whole service is self-supporting and the National Science Foundation finances research and development on the Society's information services.

This is not all. The *Chemical Abstracts* service includes specialized tertiary services such as *CA Condensates* which gives title, authors' names, journal and patent references and keyword index entries for each abstract in *Chemical Abstracts*. *Basic Journal Abstracts* gives a selection of entire abstracts. A new reference tool is ACCESS, to help people locate specific documents. It covers the chemical information in 397 libraries in 28 nations, including in the coverage 21 000 journals, monographs and conference proceedings. Dr

Cairns also described briefly the information activities of two other American societies, but not the American Physical Society.

Here we can interpolate something about the British system called INSPEC, run by the Institution of Electrical Engineers. Grants from OSTI since 1965 have enabled this service to claim to be the world's largest English-language information service, and it covers physics, electrotechnology, computers and control. It covers six abstract and current-awareness journals. Since January 1969 all its publications have been computer typeset. Research and development goes on on nine projects in the area of computerized information retrieval and dissemination. An integrated Selective Dissemination of Information (SDI) service is to become operational next January. By the end of this year there will be over 225 000 entries on magnetic tape, with an estimated 12% increase next year. An agreement has been signed between INSPEC and the Russian VINITI (of the Academy of Sciences Institute of Scientific Information) to exchange classification systems and publications and information on data-input methods.

Documentation in a learned or professional society is thus seen to be important and apart from the excellent examples given above is in need of critical examination if the dissemination of information is to be maximized. There are frequently problems of finance and there are people who clamour for government aid, but Mr Knox has emphasized that there is no reason for a society to fail to make a profit if sufficient thinking and hard work is put into an enterprise. He mentioned the American Medical Association's journal as being the most profitable magazine in the USA. However, there is obviously a vast subject here, for clearly a service may be necessary for members and for a community but not in itself be profitable. Mr Knox argued that a business that did not proide a useful service would fail to make a profit but Dr Dieter Behrens of DECHEMA (Germany) made the point that though it might be true that only a useful service could make a profit it was not true that all useful services *must* make a profit.

There is also a lot to be done about the *quality* of information-processing done by a society. The primary sources—journals or proceedings—are not read by most working scientists and engineers. As Professor Allen observed, people do not write in these journals to communicate but to gain prestige. There has to be other documentation for the majority. Dr Cairns mentioned the critical review and the state-of-the-art survey as being important and revealed that the American Society of Civil Engineers had recently instituted an annual state-of-the-art Civil Engineering Award—a plaque and an honorarium—for such reviews. Mr Larink, who obviously had given much thought to the problems of information and felt deeply about it, especially in relation to his own country of Germany, where the small total readership in German as distinct from the wide world readership in English created its own limitations, pursued the theme of readability. Without giving much detail unfortunately (for it was a fascinating point) he stated that research had shown that people could take in only three bits of information a minute. The inference from this is that a document for more general readership must have a diluent in the form of entertainment. Mr Larink claimed that it was known that engineers would accept technical literature with an entertainment component. 'It does not solve their questions but gets them to pose questions.' It is a valid point and it brings up the matter of 'popular' writing, looked on by so many scientists in Britain—but not so much in some other countries—with condescension, however friendly, patronage and even sometimes with arrogance. This is not a trivial matter—anyone who has attended learned meetings has been able to observe very eminent scientists dropping off to sleep because the discourse was too heavy going for complete absorption. And if men of such quality need a light touch so do more ordinary scientists and engineers. Societies could do worse than look at this matter seriously if they are to communicate. Admittedly my example is of the spoken word, very much more difficult to assimilate from a set lecture than the written word, but it sharpens the point.

One very interesting method of making information more easily comprehensible was described at the seminar (with illustrations from the National Institute for Medical Research and the NPL) by Mr J. A. Cade, a consultant to the Directorate for Scientific Affairs of

OECD. It was concerned with the visual presentation of scientific and technical information and the hub of his argument was that human beings, with the most remarkable computers in their heads, could absorb information by several pathways in parallel, whereas written or spoken words gave information serially. He entertained the audience with examples. He gave a serial description of a man by means of a list of characteristics and then displayed nine pictures of different men all with the characteristics given, whereas when later on he gave the full picture of the man he had poorly described serially everyone recognized it, as it were, instantaneously. He showed that facts about a patient such as temperature, pulse rate, blood pressure and so on could be given, as is commonly done, serially. A doctor trying to comprehend these facts would have to think about them in order and by means of combinations in order to understand their total significance. A graph, however, made the information much more quickly assimilable. A polar type of diagram, with the data measured along the radii and the whole completed by a continuous line through the extremities, showed the situation even more readily. A normal person would show one sort of envelope and a sick person would show another shape quite obvious at a glance. The change of such a system with time could also be shown by the use of a cathode-ray oscillograph, for coincident lines would reinforce at each presentation. A further parameter that could be used was colour. Mr Cade also demonstrated that the human brain had an extraordinary ability to reject 'noise' in the presentation of data. The whole purpose of his discourse was to demonstrate the power of the brain as a data-processing system, much greater than that of any electronic computer, and so indicate how scientific and technical information could be presented for ready assimilation in a way quite different from that of the spoken or written word.

These few extracts from the proceedings of a seminar which was not public and was intended for OECD use, with my own additions and reflections, may serve to get physicists thinking about the changes in business organization, documentation, conference and meeting arrangements and presentation that would help deal more satisfactorily with that essential commodity, information. It is essential to all science and technology but more pressingly to the use of those in industry.

# Book reviews

***Theory of Elasticity***, by L. D. LANDAU and E. M. LIFSHITZ. (Oxford: Pergamon Press, 1970.) Pp. viii+165. £3.

It is almost presumptuous to review a book by such eminent physicists as L. D. Landau and E. M. Lifshitz, especially when the first edition was generally acclaimed. However, some indication of the scope and treatment of the subject will be useful to those coming fresh to 'Landau and Lifshitz'. The first chapter deals with fundamental equations. It uses tensor notation from the start, it is terse and economical, and seems to me to give a clearer idea of the basic notions than some more conventional treatments. I do have however one reservation here and throughout the book—I cannot see the point of using bulk modules and rigidity or Young's modules and Poisson's ratio, instead of the Lamé parameters, $\lambda$ and $\mu$. The latter almost always lead to simpler expressions, especially for the wave equations, and it is the generalized Lamé parameters that must be used in dealing with anisotropic media. While the mathematics is economic and elegant, the emphasis is on the physics of the phenomena discussed, which comprise the equations of equilibrium, especially for plates and bars, wave motion (a particularly clear and neat discussion), vibrations, dislocations, and thermal effects.

The English translation is a pleasure to read, entirely clear, well printed and with good diagrams.

A. H. COOK

***Methods of Experimental Physics, Vol. 8. Problems and Solutions for Students*** edited by L. MARTON and W. F. HORNYAK. (London: Academic Press, 1969.) Pp. xii+275. £6 15*s*. (£6.75).

The intention of this book is to provide graduate students with problems concerning the general design considerations of experiments, and this is meant to include such topics as signal-to-noise ratios and the feasibility, limits of accuracy, and precision of experiments. Although this is undoubtedly a very worthwhile intention, the present book has only a limited degree of success in achieving it. This is especially unfortunate since the other volumes in this series have been outstandingly good.

The real weakness of the book is that many of the chapters are so highly specialized that they are of use only to students working in these particular fields. There is a need for a book of this type which gives information and practice on experimental problems which are common to a wide range of different experimental fields, such as fundamental limitations of accuracy, control of conditions by feedback devices, numerical techniques in data analysis, and computer control of experiments. Some general topics such as these are dealt with in the book, but not enough to make it a book which which one could recommend to graduate students for general reading.

F. H. READ

***MHD Power Generation: Engineering Aspects***, by G. J. WOMACK. (London: Chapman and Hall, 1969.) Pp. x+155. £3.

Magnetohydrodynamic (MHD) power generation cuts right across the stereotype compartments in English engineering education, a considerable disadvantage for any large project.

This small book is a very readable introduction to the field and is directed mainly to ideas for open cycle fossil fuelled plant. It has five chapters ranging through electrical conductivity of gases, electrodynamics, MHD duct geometry, power generation cycles, to components of open cycle MHD plant.

As each new topic is introduced the author endeavours to start at an elementary level to accommodate the wide range of backgrounds expected in its readers. Each chapter is well provided with graphs and drawings and a useful list of references.

There is no discussion of MHD in its economic context.

P. J. FOSTER

***Propagation of Waves***, by P. DAVID and J. VOGE, translated by J. B. ARTHUR. (Oxford: Pergamon Press, 1969.) Pp. x+329. £4 5*s*. (£4.25).

This reviewer enjoys generality and therefore opened this new book on the 'Propagation of Waves' with hopeful anticipation of good things to come. These hopes were soon dashed. 'Waves' turns out to mean radio waves and the 'Propagation' characteristics treated are those involved in terrestrial radio links. Final despair followed from the further discovery that the subject is treated in the traditional style of an engineering handbook. There is much in the way of facts, formulae and figures but little to be gained in the way of physical understanding.

The book is not intended to be read. It is a practical handbook for radio communications engineers. This was the authors' intention and they have succeeded in packing a tremendous amount of information into a relatively slim volume. The first three chapters cover the general features of antennas, transmission media, reflection and refraction. This is followed by 150 pages on the rôles of the terrain, the troposphere and the ionosphere. Then, after a brief chapter on interference, the final 100 pages are devoted to a detailed discussion of propagation characteristics at wavelengths from 2 km to 1 cm. The complete absence of vector notation gives the text an archaic appearance, but it is always sound and makes up in detail what it lacks in depth.

The book provides a useful source of data about terrestrial radio links. It is a pity that the title gives so little indication of the true nature of the contents.

P. N. BUTCHER

***Tensor Properties of Materials***, by A. R. BILLINGS. (Chichester: John Wiley, 1969.) Pp. xv + 171. £3 10*s*. (£3.50).

The subject of crystal properties has been well covered by text books beginning with the classic by Voigt and followed by Wooster, Nye, Ramachandran and Mason among others. Compared with Nye's book which is a model for clarity, this book is more condensed but is more up to date and places more emphasis on transport properties and discusses time inversion invariance and its effect on compliance and rigidity tensors.

The first chapter is a useful though condensed treatment of thermostatics. It suffers perhaps from trying to introduce provisionally subjects like suffix notation, tensors, stress and strain. The second chapter deals with the definition of tensors and introduces generalized compliance and rigidity. The next three chapters deal with crystal periodicity, space and point symmetry transformations, the 32 crystal classes, Neumann's principle, pseudotensors and time inversion and its influence on property tensors. The province of the last three chapters is the transport of heat and electricity in homogeneous and inhomogeneous solids both in the absence and in the presence of a magnetic field. Onsager's Principle is used and the idea of a generalized resistivity introduced.

There are seven appendices including an extensive one on the crystal classes which has stereograms of the 90 magneto-crystalline classes in which those cases including time inversion symmetry are depicted in red colour.

One partial omission noted is a thorough discussion of non-linear properties which are now of increasing importance. In this context it might have been useful to make more

distinct the difference between the partial derivatives making up the generalized compliances and the linear approximations to them, known to us more familiarly as permittivity permeability, etc.

E. L. JAMES

***An Introduction to Photobiology*** by YVES LE GRAND. (London: Faber and Faber, 1970.) Pp. vii + 178. £2 10*s* (£2.50).

This book concerns a field of scientific activity of recent popularity but not of recent origin. The effects of sunlight on man and other biological material have been studied for more than a century. During the last thirty years, however, there has been an upsurge of interest in the interactions of photons with a variety of biological objects and in the use of the photon as a very sensitive tool in biological research. Photobiology is one of the 'frontier sciences' of today, calling for the contributions of chemists, physicists and biologists before any of its phenomena can adequately be studied.

Professor le Grand's main interests lie in the fields of vision, colour and problems of illumination and about one half of his book deals with photoreception in animals and various other topics such as visual pigments, animal responses and colour vision. His introduction to these subjects is concise and clear. Earlier chapters on theory and techniques used in photobiology are more brief and probably most useful to the biologist setting out on photobiological research. A chapter on 'Ultraviolet and the living cell' deals with harmful and mutagenic effects of ultraviolet radiation, photodynamic action, and that popular topic, photoreactivation, first described in 1949 by Kelner when he observed repair of the lethal effects of ultraviolet by subsequent or simultaneous irradiation with visible light. A later chapter describes effects of light on the skin, animal photoperiodism, as in seasonal rhythms in reproduction and Circadian rhythms, those recently described (1959) 24-hour periodicities of various biological effects. Bioluminescence is treated in the final chapter.

This book does not discuss all the topics usually included under the umbrella of photobiology. Photosynthesis, one of the most important of these, is not mentioned nor are many of the effects of light on man. However, those subjects which are summarized are treated in a manner which should stimulate the interest of the physicist or chemist knowing little biology and looking for new fields to conquer, or of the biologist who recognizes the fascination of these biological phenomena but needs guidance in the theories and techniques to be mastered for these investigations.

The author's frequent reference to photobiological problems awaiting study and to our limited understanding of many others should stimulate younger scientists to enter this expanding field of biological research.

One error in translation was noted: on p. 146, for *sunstroke*, read *sunburn*.

E. M. F. ROE

# Book notices

***Personnel Dosimetry Systems for External Radiation Exposures***, Technical Reports Series No. 109. (Vienna: International Atomic Energy Agency, 1970.) Pp. 150. £1.67.

This guidebook, which has been prepared with the aid of a group of consultants, describes the characteristics, applications, practical advantages and limitations of various types of dosimetric systems that are being used at the present time for the assessment of external radiation doses.

***Experimental Crystal Physics***, 2nd edition by W. A. WOOSTER and A. BRETON, (London: Oxford University Press, 1970.) Pp. ix+144. £6.

This edition continues the tradition of the first edition in presenting experiments that can be performed in two or three hours. The additional subjects covered include: dielectric constants, dynamic measurements of piezo-electricity and elasticity, the growth of crystals, semiconductors, the Hall effect, ferroelectricity and ferrimagnetism.

***Techniques and Methods of Polymer Evaluation***, Vol. 2, ***Thermal Characterization Techniques***, edited by P. E. SLADE, JR and L. T. JENKINS. (New York: Marcel Dekker, 1970.) Pp. x+371. £8.90.

This book is concerned with a number of techniques that are of more specialized interest than the broad treatment of differential thermal analysis and thermogravimetric analysis contained in Volume 1 of the series. These include: differential scanning calorimetry, pyrolysis–gas chromatographic techniques, torsional braid analysis and electrothermal analysis of polymers.

***Glass Lasers***, by K. PÁTEK, edited by J. G. EDWARDS. (London: Butterworth, 1970.) Pp. xi+217. £6.50.

This book discusses solid-state lasers characterized by a resonant cavity with a glass of crystal active element pumped by a light source. It was written by Dr Pátek before his death in 1967 and the editor has updated the text and added a representative bibliography to cover the most important developments since the book was written.

***Electrical and Magnetic Ceramics* (2), *Proceedings of British Ceramic Society***, No. 18. (Stoke-on-Trent: British Ceramic Society, 1970.) Pp. 303. £5.25.

This book contains papers presented at a meeting of the Basic Science Section of the Society held at the University of Warwick on 9–10 September 1969.